THE PARIS JOURNAL ON AI & DIGITAL ETHICS

Editorial Committee

Dr. Thomas Souverain, Editor-in-Chief (French Alternative Energies and Atomic Energy Commission)

Dr. Hubert Etienne, Editor (Quintessence AI)

Dr. Herbert Lin, Thematic Chair (Stanford University)

Prof. Joshua Tucker, Thematic Chair (New York University)

Prof. Aggelos Kiayias, Thematic Chair (University of Edinburgh)

Dr. Alexei Grinbaum, Thematic Chair (French Alternative Energies and Atomic Energy Commission)

November 2025

A New Journal for the 2025 Proceedings of the Paris Conference on AI & Digital Ethics

Entering a digital era in which opacity arises not only from scarcity but also from the *overabundance* of data, the time is more than ripe to summon philosophy. In the fifth century BCE, Socrates devised the *maieutic* method in order to give birth to truth despite the plurality of opinions.[1] As for us, in front of contents of social networks or platform discussions where recommendations, predictions, or generations of Artificial Intelligence (AI) models come into play, we must learn to discern the authentic from the fake, and the beneficial from the harmful.

In this spirit, we present the **Paris Journal on AI & Digital Ethics (PJAIDE)**. The journal brings together the 20 full papers of abstracts presented at the *Paris Conference on AI & Digital Ethics*,[2] held for the fourth year at Sorbonne University on 16–17 June 2025. Like the Paris Conference, the journal unites the efforts of scholars in humanities, social sciences, but also cognitive and computer science, in developing critical and constructive approaches to the challenges raised by AI systems and digital transformation. The current edition is built around four tracks:

I. Controlling Cyber-Influence and Manipulation Strategies examines the issue of disinformation and misinformation in all its complexity ;

II. Countering Information Manipulation Techniques introduces and analyses advanced strategies for addressing these challenges ;

III. Exploring the Potential of Blockchain to Renew Civic Trust investigates how blockchain and cryptography can be leveraged to establish trust in the exchange of AI-generated content ;

IV. Rebuilding Social Cohesion in AI-Powered Societies expands the discussion to broader, long-term questions of trust, examining both the moral desirability of current AI applications and their potential to serve civic purposes.

These tracks address their two central dimensions — information and trust — through scholarly rigour and a shared commitment to directing technological progress towards the common good. The quality of the contributions reflects the work of the thematic chairs — Herbert Lin, Joshua Tucker, Zeve Sanderson, Alexei Grinbaum, and Aggelos Kiayias — and of the anonymous double-blind reviewers, to whom I express my gratitude. My final thanks go to my philosophical colleague, Hubert Etienne, who founded the remarkable Paris Conference and entrusted me with the honour and responsibility of leading the scientific process.

I warmly encourage all inquisitive minds who are not afraid in going beyond the frontiers of their own discipline, in coupling blossomed academic arguments and nitty-gritty technical details, in fusing intellectual honesty with civic commitment, to engage with the pages of PJAIDE.

Thomas Souverain, Editor-in-Chief

thomas.souverain@cea.fr

[1] See for instance Plato, *Theaetetus*, 148 d - 148 e 7 ; 157 c 7 - 157 d 5.

[2] For more information on the yearly conference and calls for contributions, please refer to the website of PCAIDE: https://paris-conference.com/.

Contents

I. CONTROLLING CYBER-INFLUENCE AND MANIPULATION STRATEGIES .. **4**

The Landscape of Information Dysfunction: A Visitor's Guide **4**

Herbert Lin.....

Digital Curtains Descending: Measuring and Comparing Legal Closure over Digital Spheres after TikTok v. Garland **13**

Davide Clementi.....

The Demand for Bullshit under American and European Law..... **27**

Lila Greenberg, Julie Wittenberg, Jane Bambauer .....

From Disinformation Dissection to Ideological Conflation: Leveraging Semantic Embeddings and Stochastic Processes for Detection of Behavioral Drifts Among Content-Creators **38**

Clément Benesse.....

Who Controls the Narrative? The Dual Role of Contributors and Amplifiers in Online Activism..... **46**

Soo Young Bae, David DeFranza.....

II. COUNTERING INFORMATION MANIPULATION TECHNIQUES..... **58**

Introduction on countering information manipulation..... **58**

Joshua Tucker, Zeve Sanderson .....

Gaming the System: Extremist Infiltration Strategies in MMORPGs and the Metaverse **63**

Didier Danet.....

A Socioinformatic Analysis of Community-based Fact-Checking: Why Community Notes is Not as Effective as X Claims **76**

Julian M. Pelloth, Lena Pölzer, Katharina A. Zweig.....

Speech Act Gaps and Counter-Speech **96**

Isaac Taylor.....

Accountability and Attribution in AI-Generated Content Authentication: Lessons from China's AI Content Labelling Mandate..... **105**

Wenlong Li.....

III. EXPLORING THE POTENTIAL OF BLOCKCHAIN TO RENEW CIVIC TRUST **120**

Trust By Decentralization? Where Blockchain Systems Meet Responsible AI Challenges..... **120**

Aggelos Kiayias .....

Building Trust and Developing Value between Creative Rightsholders and GenAI through Tokenised Licensing and Content Authenticity Tools..... **126**

Chris Elsden, Frances Liddell, Caterina Moruzzi, Burkhard Schafer, Ella Tallyn, Evan Morgan, Billy Dixon, Kar Balan, John Collomosse.....

From Black Box to Glass Box: Leveraging Blockchain to Audit AI Systems Through Multistakeholder Participation..... **140**

Roberto Caparroz de Almeida, Lucas Maldono Latini .....

Bootstrapping Trust across Web2 and Web3 Domains Using Publicly Verifiable Web Data **153**

Yuan Lu, Qiang Tang .....

IV. REBUILDING SOCIAL COHESION IN AI-POWERED SOCIETIES **165**

Rebuilding Social Cohesion: AI & The Faces of Trust **165**

Thomas Souverain .....

Who Owns the Words? Copyright, AI Prompts, and the Governance of Digital Authorship **171**

Jaqueline Simas C. de Oliveira .....

Critical Thinking and Media Literacy in Youth from Underserved Contexts: Evidence from a Field Intervention **181**

Alice Bougnères, Hugo Besançon, Melisa Basol .....

AI, Digital Democracy, and the Crisis of Social Cohesion: Rebuilding Political Trust **196**

Adrien Tallent .....

AI as Bureaucracy: The Standardisation of Judgment **211**

Carina Prunkl .....

The Trust Paradox: How Companion AI Is Rewiring Human Connection and Social Cohesion **213**

Alva Markelius, Priscila Chaves, Sarah W. Spencer, Joahna Kuiper .....

AI with Heart: A Joy-Centered Framework for Ethical AI Development & Implementation **231**

Desmond Upton Patton, Shana Kleiner, Shug Miller, Nicholas Deas, Jessica A. Grieser, James Shepard, Blake Vente, Kathleen McKeown .....

The Landscape of Information Dysfunction: A Visitor's Guide

Herbert Lin

Department of Cyber Policy and Security, Hoover Institution, Stanford University.

Corresponding author. E-mail: herblin@stanford.edu

Abstract

Modern society faces a rapid increase in the volume, speed, and complexity of information flows. This overview outlines the information environment's technological, informational, and cognitive dimensions; the principal actors that shape it; and the types of communicative activities and dysfunction that emerge. Beyond mis-, dis-, and mal-information, we catalog additional problematic forms and describe actor-specific drivers of dysfunction. We consider how platform incentives, citizen cognition, and government behavior interact with accelerating technologies to degrade deliberation, and we situate four companion papers within this broader landscape.

Keywords

Information Environment, Misinformation, Disinformation, Malinformation, Platform Incentives, Polarization, Media Ecosystems, Digital Manipulation.

1 Introduction

Modern society in the early twenty-first century has experienced a significant increase in the volume, speed, and complexity of information that circulates within it as compared to decades past. The digital revolution has changed how information is produced, distributed, consumed, and used. Social media and artificial intelligence as they are integrated into the daily information habits of the body politic are much discussed—and rightly so—as important factors in what is increasingly described by parties of various political persuasions as a dysfunctional information environment. But they are not the only or even necessarily the most significant factors.

The four papers of this section address techniques and strategies for cyber-influence and manipulation of information. To frame their contributions, this introduction surveys the broader landscape of information dysfunction. It does not propose remedies—that is the focus of the papers themselves—but instead outlines the diverse forms of dysfunction that undermine the integrity and reliability of information ecosystems, highlighting the complexity of the challenges these works address.

To state the bottom line up front—dysfunction in the information environment is not simply a function of mis-, dis-, and mal-information. Many other types of information that do not conform to the strict definitions of these acknowledgedly important types of information are important as well… and all depend strongly on the context of expression.

2 The Information Environment

The information environment has an intuitive meaning for most people, but it is worth exploring in more detail. The central construct underpinning this visitor's guide is the information environment (IE)(US Defense, 2023). This environment encompasses all the social, cultural, linguistic, psychological, technical, and physical elements that shape how both humans and automated systems derive meaning from, act upon, and are impacted by information. The IE is not static or monolithic; it shifts in response to technological change, behavioral trends, and societal context. Understanding its structure helps clarify the nature and severity of information dysfunction. The IE has three primary dimensions:

- **The technological/physical dimension encompasses the hardware and software** that facilitate the creation, storage, transmission, and processing of data and messages. The infrastructure—ranging from personal devices to global fiber networks and the algorithms that govern platform behavior—serves as the physical backbone of the information space. It sets the parameters for how quickly, efficiently, and widely information can be created, modified, and accessed.
- **The informational dimension is the content itself** —regardless of whether it is text, audio, video, images, data tables, or any other form. It includes every aspect of collection, organization, processing, storage, protection, and dissemination. Informational content is deeply shaped by how it is structured and delivered (e.g., via articles, tweets, podcasts, or emails), and by the ease or difficulty involved in repurposing or remixing content across platforms and media.
- **The cognitive dimension** covers the attitudes, beliefs, values, perceptions, internalized habits, and emotional responses of people as they receive, interpret, or respond to information. Cognition is filtered through individual life experience, culture, history, and social context. As a result, two people (or audiences) may interpret the same material in radically different ways, depending on their previous beliefs and experiences.

These dimensions intersect continuously. For example, a technological advance, such as easy-to-use video editing tools, may radically change the kinds of informational content produced and, in turn, the way audiences interpret what they see or hear.

A key idea about information environments is that information almost never moves from creator to receiver directly. Instead, it traverses networks comprised of intermediary nodes—editors, platforms, reposters, and consumers. Each step increases the chance for both inadvertent and deliberate distortion. When information passes through deep or complicated networks, the original context, meaning, and even intention may be altered or lost, sometimes drastically. This makes approaches to cleaning up disrupted information environments difficult, since problematic media may propagate faster than can be controlled(Vosoughi, 2018).

The above thumbnail characrerizaton of the information environment can be positioned in the context of the academic literature, the best example of which is a comparble and more elaborated view of the information environment is provided by Alicia Wanless, who uses ecological metaphor for the information

environment to conceptualize it as a complex, adaptive ecosystem composed of humans, technology, and information interacting dynamically(Wanless, 2025).

In particular, the information environment is a complex system with many interacting agents (humans, technologies, machines) shaping and being shaped by the flows of information. Like a natural ecosystem, it experiences disturbances (e.g. misinformation campaigns, viral content floods) that compete for dominance and may cause "information pollution" - what this introduction calls dysfunction - in the form of spam, infotainment, deliberate lies, or manipulation. Competing and interacting ideas analogous to species competing for resources or niches or influence, such as attention or political groups or political power.

3 Actors in the Information Environment

As noted above, the information environment is highly dynamic. The agents of this dynamism are diverse, and their behaviors and decisions shape the state and health of the ecosystem as each of them receives from the environment, processes that shape and re-shape information locally, and then transmit it back into the environment. The scope, nature, and extent of their information-shaping activities are governed by a varying mix of laws, regulations, norms of behavior, policies, personal preferences, politics, cultural and social influences, psychological factors, economic conditions, and environmental contexts. Technology acts as an overlay on all of these factors.These agents (actors) include:

- **Citizens** are the body politic—ordinary people who shape demand for information, participating in sharing, and thereby molding public discourse.
- **Journalists** are responsible for gathering, verifying, and presenting news and analysis, journalists function as arbiters and curators of facts, giving context and scrutiny to fast-moving or contentious topics. They also play important roles in agenda setting and directing attention.
- **Universities and research institutions** are sources of new knowledge and providers of education, these institutions promote methods for gathering evidence and exercising critical thinking. They also serve as repositories for evidence-based research and analysis.
- **Libraries** play a traditional but essential role as providers of access to reliable information and promoters of information literacy. Libraries' stewardship of archives, books, and digital materials supports public and academic inquiry.
- **Media** organizations include print, broadcast, and digital media. These entities distribute news, analysis, and opinion, shaping what narratives reach and influence broad audiences. Their decisions about coverage have powerful ripple effects on society's sense of priorities and understanding.
- **Technology providers** are platforms such as social media sites, search engines, and infrastructure companies that host, index, and algorithmically moderate huge amounts of information, often at a scale never before seen. Their design choices, terms of service, and content-moderation practices have a deep impact on what information surfaces or circulates.
- **Political parties and centers of influence** help to frame debates, set agendas, and organize engagement in their efforts to advocate for specific policy outcomes, to mobilize populations, and to influence the flow and context of information.
- **Governments** set regulations, maintain legal frameworks, and are significant sources of information and data themselves. Governments may act as both information providers and regulators, supporting or undermining transparency and accountability based on prevailing political or strategic imperatives.

4 Types of and Venues for Communicative Activity

Actors in the information environment engage in a diverse range of communicative activities. These activities, described below, are not necessarily used or selected with conscious intent to suit specific audiences, contexts, and strategic objectives, but observed from a distance they can often be seen as having effects as indicated below. In practice, different types of activity are often used together in what appears to be—and indeed well be—deliberate attempts to tailor, amplify, and reinforce information across platforms to influence perception, shape discourse, and achieve desired outcomes. Five basic types of communicative activity are prominent:

- **Informational activities** provide new or clarifying information.

- **Influence activities** seek to persuade or shift beliefs, attitudes, or behaviors of others.
- **Attention manipulation activities** are designed to attract, direct, or divert attention, sometimes prioritizing certain topics over others.
- **Relational/evocative activities** build relationships, affirm identities, or evoke emotional responses.
- **Entertainment activities** provide amusement and pleasure for citizens. Such activities may blend or mask elements of the other four.

These activities are rarely isolated; most real-world communications blend aspects of multiple types. For instance, decision-making often integrates informative, persuasive, and evocative elements, while public safety campaigns might simultaneously seek to inform, influence, and reassure the public. Actors operate these activities across a variety of venues, including:

- **Social media** (Facebook, Twitter/X, TikTok, Instagram);
- **Messaging systems** (email, SMS, WhatsApp);
- **Broadcast and print media** (TV, radio, newspapers, digital news);
- **Specialist media** (journals, industry newsletters);
- **Public events** (rallies, town halls, church services);
- **Formal and informal education**;
- **Art, advertisements, and even graffiti**.

Any communication in these venues can transform the IE, subtly or dramatically shifting collective awareness, values, and priorities.

5 Types of Potentially Dysfunctional Information

Discussions about dysfunction in the IE center on three main harmful types: misinformation, disinformation, and malinformation(Wardle and Derakhshan, 2017). These are an important point of departure for understanding the IE, but a description of information dysfunction cannot end there.

- **Misinformation** is false or misleading information shared without the intent to deceive. Misinformation often results from honest errors, misinterpretation, or poor verification, and is frequently spread by individuals who believe it is true.
- **Disinformation** is deliberately fabricated or manipulated information, crafted to deceive, mislead, or cause harm. Disinformation encompasses everything from conspiracy theories, deliberately staged images, forged documents, and coordinated propaganda campaigns. Even when passed on by parties without deceptive intent, it becomes misinformation.
- **Malinformation** is factual information broadcast maliciously, often stripped of appropriate context, to hurt individuals or institutions (e.g., doxing, leaking classified documents, or selectively quoting true statements for misleading effect).

Because of the specificity of their definitions, these categories do not cover the space of possible problematic forms of ”untrue” or misleading information, many of which defy simple categorization. For example, consider:

- **Fiction, satire, and parody** are not inherently deceptive, but they may be misinterpreted as real, especially when circulated beyond their original context. Examples include satirical headlines from “The Onion” mistaken for real news, or “deepfakes” portraying realistic but entirely fictional audio-visual events.
- **Sarcasm** depends on tone, nuance, and context that are often lost in pure text renditions, leading to confusion. Sarcasm is also plausibly deniable if a statement prompts backlash.
- **Curation/omission** is almost always unacknowledged. When news outlets or platforms consistently prioritize or suppress coverage of certain stories, they shape public consciousness, sometimes in ways that benefit entrenched power or particular worldviews.

- **Dog whistles and coded language** are examples of ambiguous language whose use signals deeper meanings to in-groups (e.g., "West Korea" as code for criticism of China), bypassing censorship or plausible deniability.
- **Non-mentioning mention/deflection** involves raising an issue by pretending not to raise it, or hinting at conspiracy with statements such as, "I don't bring this up, but others are talking about it."
- **Blurring fact and opinion** occurs when editorials are presented as news, or commentary hides in the format of hard reporting. Such blurring is intended to muddle boundaries and confuse audiences, and generally weakens trust in journalism.
- **Leading questions and false/misleading provenance** imply fact and grant unwarranted authority or credibility.

6 Dysfunctional Communicative Activity

All IE actors sometimes engage in behaviors that drive dysfunction. They may be deliberately trying to promote dysfunction or they may be doing so without any conscious intent to do so, even though that may be the effect. Illustrative examples follow below.

- **Dysfunctional informational activities** are often the result of time pressures. Pressure for rapid response in crises leads to premature or inaccurate reporting, and corrections seldom erase the impact of initial errors. Journalistic haste and democratized publishing platforms sidestep traditional vetting, fact-checking, and peer review(Ross Arguedas, Badrinathan, and Mont'Alverne, 2022). Engagement-optimized algorithms deliver highly arousing but low-fidelity content, often crowding out reliability. Mistrust in traditional gatekeepers further reduces reliance on reputable institutions. Over time, these forces lead to information silos and echo chambers.
- **Dysfunctional influence activities** differ from persuasion because they manipulate, deceive, or intimidate. Peer pressure can enforce homogeneity, stifling dissent or skepticism. Appeals to authority or to personal charisma substitute structure for substance. Appeals to anger and fear drive emotional decision-making and polarize groups. Simple narratives and memes are used to decontextualize questionable arguments, making them more palatable and persistent even when wrong(Masood, 2024).
- **Dysfunctional attention-manipulating activities** prioritize interests other than those of the audience. Distraction, such as "whataboutism," draws attention from uncomfortable or damaging issues. Agenda-setting controls what the public considers important by highlighting select stories. Deliberate overload and noise generation fragment attention and reduce public capacity for considered action.
- **Dysfunctional relational/evocative activities** can often be inferred when activities provoking outrage deepen group identity and amplify polarization. Outrage is engineered to elicit instant, uncritical responses, making rational deliberation or compromise difficult. Group identity is repeatedly affirmed at the expense of cross-group trust, producing echo chambers and mutations of confirmation bias.
- **Dysfunctional entertainment activities** such as "infotainment" and drama-driven media blend information with engaging content (e.g., dramatic visuals or exaggerated headlines), often sacrificing factual accuracy and critical evaluation for audience attention and emotional impact. Fictional narratives can influence belief, as shown in cases like *Wag the Dog* (a movie in which a US president starts a war to gain political advantage). Repeated exposure to outlandish narratives on a wide scale can normalize fringe opinions or behaviors(Bolet and Foos, 2025).

Small amounts of these activities are tolerable or even necessary in society, but at scale, they undermine collective understanding and norms of rational behavior.

7 Actor-specific Examples of Dysfunctional Information Activity

How would one recognize dysfunctional information activity? It is helpful to describe some of the observable (or at least inferable) characteristics of such activity organized by actor.

7.1 Dysfunctional Communicative Activity of Platform Providers

Making money by providing plattorms that facilitate communication depends on one (or both of) two methods: charging the communicating parties directly for the services they receive or selling ads to third parties with something to sell and encouraging or forcing communicating parties to watch them. The latter is most common in today's information environment and it shapes the business models of for-profit companies.

Specifically, these business models have distinct financial incentives to maximize the number of ads that they can show to consumers, and thus to maximize engagement and dwell time in using these platforms. Thus, one might expect to see user engagement prioritized over other values, such as providing accurate and properly contextualized information or promoting social cohesion. Algorithmic amplification thus promotes sensationalism over truth. Content curation, based primarily on inferred user preferences, erodes a common societal knowledge base by creating personalized echo chambers(Bakshy, 2015). Easy sharing mechanisms facilitate impulsive, low-effort spread of questionable content, stretching these echo chambers to networks of users who share similar ideas, creating information silos(Denniss and Lindberg, 2025). Opaque algorithms, sponsorship, and moderation reduce transparency and accountability.

7.2 Dysfunction in Communicative Activities of Citizens

Many citizens lack critical media and information literacy, limiting their ability to judge sources or claims quickly or effectively(Ecker, 2022). Emotional or identity-driven misinformation is often valued more than factual accuracy. Likewise, citizens rarely have personal incentives to recognize irrelevant or non-factual information(Downs, 1957). They are highly motivated to want information that will help them to make better purchasing decisions, to help them in doing their jobs, and to find pleasing entertainment. But in their roles as voters and citizens, their desire to be "good citizens" must be placed against the pressures from of their social groups, families, and friends to "vote the right way."

Human limitations in cognitive structure defaults to rapid, intuitive thinking (System 1) over slow, reasoned processing (System 2). Affective and novelty biases arising from quick heuristics further distort perceptions, making emotionally resonant or simply new information more compelling, regardless of accuracy. Authenticity is often valued over expertise, reducing uptake of well-informed perspectives.

Finally, the sheer amount of information available to individuals contributes to information dysfunction for two main reasons. First, many individuals lack the critical skills needed to assess the credibility of the information they encounter, making it difficult to distinguish reliable sources from misinformation. Second, the vast volume of information enables people to selectively find support for any preconceived or unfounded belief, reinforcing biases and spreading false or extreme ideas.

7.3 Dysfunction in Communicative Activities of Governments

In some nations, governments assume some responsibility for promoting a healthy and functional information environment. Others do not, and indeed they may take deliberate actions that deform and distort environment. They have a variety of approaches for doing so.

One of the most obvious is found in officials and agencies that themselves promote misinformation and malinformation, endorsing false narratives such as election result denials, manipulated statistics, or fabricated crises to justify political aims. These actions normalize disinformation as a governing tool and erode public trust in truth and institutions.

Suppression and manipulation of information have likewise intensified. Officials pressure media to limit critical coverage both directly and indirectly (the latter by threatening to exercise discretionary authority against media entities over unrelated areas), demand pre-approval of scientific findings, or restrict access to data. Watchdogs are dismissed, and informants are recruited to monitor dissent. Citizens face surveillance, including scrutiny of social media at border crossings, for expressing critical views.

Governments also seek control over independent knowledge institutions. For example, they may seek to place universities under receivership(Gluckman, 2025), withdraw previously approved peer-reviewed funding, or disparage resident expertise. During the COVID-19 pandemic, officials who contradicted political narratives were publicly attacked, illustrating this disdain for professional authority(Forgey, 2020).

Finally, the quality of official communication declines as statements grow inaccurate(McFall, 2025), careless, or overtly hostile(not, 2021). Sloppiness in the phrasing and presentation of such documents is notable when similar documents, regardless of content, from previous administrations have been carefully scrubbed for proper language and grammar.

8 The Impact of Technology on the Information Environment

In 1927, Justice Louis Brandeis of the US Supreme Court stated that:

"No danger flowing from speech can be deemed clear and present, unless the incidence of the evil apprehended is so imminent that it may befall before there is opportunity for full discussion. If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence. Only an emergency can justify repression."(Court, 1927)

Notice that Justice Brandeis' prescription both the existence of a healthy "marketplace of ideas" and sufficient time for discussion and correction. Today, however, the massive increase in the volume and velocity of information, supercharged by social media and automation, means that time and opportunity for critical correction are both in short supply.

Today's technologies now offer high connectivity, anonymity, customizable searches, borderless reach, and cheap, accessible publishing tools. This enables not only large-scale, coordinated information campaigns (both benign and harmful) but also the amplification of rumors, leaks, and propaganda by individuals and groups who would have lacked any reach even a generation ago. Powerful actors can exploit bot networks and targeted advertising to spread messages with previously unimaginable efficiency and subtlety.

9 The Papers of This Session

With the preceding as background, we now turn the papers of this session on controlling cyber-influence and manipulation strategies. Digital Curtains Descending? notes that different regulatory architectures—nominally intended to protect national publics from pernicious cyber-influence and manipulation—emerge from fundamentally different national conceptions of state authority, and that this is true even amongst nations that appear to be like-minded. Ambitiously, it asserts the realistic possibility that these regulatory architectures could converge in a manner consistent with international cooperation on such matters.

The Demand for Bullshit Under American and European Law addresses the demand side of a dysfunctional information environment, noting that most interventions to control cyber-influence and manipulation strategies try to choke off the influence of active contributors of dysfunction. The novel approach of this paper is to proposes policies that treat media as a vice product—one that is inherent detrimental to society independently of the information content they carry—and thus that can be regulated accordingly.

From disinformation dissection to ideological conflation posits that slow co-optations of trusted information sources (e.g., content creators and influencers) that gradually shift their ideological tilts are difficult to detect and are therefore dangerous. The authors therefore propose a specific mathematical approach to characterize the ideological orientations of creators and influencers from their online footprints, thereby allowing well-known "out-of-distribution" detection techniques to identify subsequent changes in ideology and thus to signal attempts at co-optation.

Who Controls the Narrative? examines the influence of two types of actor in the information environment: content contributors and content amplifiers. The former introduce a wide variety of different perspectives, but it is the choices that the latter make regarding which perspectives to amplify that shapes the long-term visibility and framing of activist discourse in the IE. This insight suggests that interventions to control cyber-influence and manipulation strategies may be more productive when focused on amplifier networks.

The four papers share two key themes on cyber-influence. First, they recognize that regulation is complex, highlighting divergent national frameworks that challenge cooperation. Such divergences will mean different kinds of interventions for suppressing active contributors to information dysfunction and the amplifier networks that shape narratives. Second, the papers offer a nuanced view of information

actors, proposing innovative detection methods and underscoring the need to address societal vulnerabilities beyond just blocking harmful content. These insights will ultimately inform a more precise, adaptive approaches to managing cyber-influence and manipulation.

Acknowledgements

The assistance of Eric Heng in providing research support for this piece is gratefully acknowledged. Of course, any errors are mine alone.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

(2021) New conflicts in the briefing room: Using sentiment analysis to evaluate administration-press relations from clinton through trump. Political Communication 38 pp 241—-59. URL https://doi.org/10.1080/10584609.2020.1763527

Bakshy E (2015) Exposure to ideologically diverse news and opinion on facebook. Science 348 pp 1130—-32. URL https://doi.org/10.1126/science.aaa1160

Bolet D, Foos FB (2025) Media platforming and the normalisation of extreme right views. ritish Journal of Political Science 55 URL https://doi.org/10.1017/S0007123425000195

Court UJS (1927) Whitney v. california, 274 u.s. 357. URL https://supreme.justia.com/cases/federal/us/274/357/

Denniss E, Lindberg R (2025) Social media and the spread of misinformation: Infectious and a threat to public health. Health Promotion International 40, no 2 pp 1—-10. URL https://doi.org/10.1093/heapro/daaf023

Downs A (1957) An Economic Theory of Democracy. New York: Harper & Row

Ecker U (2022) The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology 1, no 1 pp 13—-29. URL https://doi.org/10.1038/s44159-021-00006-y

Forgey Q (2020) 'everyone is lying': Trump undercuts public health officials in fresh attacks - politico. URL https://www.politico.com/news/2020/07/13/trump-questions-public-health-experts-twitter-359388.

Gluckman N (2025) The trump administration told columbia to put a department in 'academic receivership.' what is that? The Chronicle of Higher Education URL https://www.chronicle.com/article/the-trump-administration-told-columbia-to-put-a-department-in-academic-receivership-what-is-that.

Masood M (2024) Political meme use can lead to political intolerance: Evidence from a panel study. International Journal of Public Opinion Research 36, no 4 URL https://doi.org/10.1093/ijpor/edae052

McFall MR (2025) Trump colombia tariff announcement had major spelling mistake - newsweek. Newsweek URL https://www.newsweek.com/donald-trump-colombia-tariff-announcement-spelling-mistake-2021167.

Ross Arguedas A, Badrinathan S, Mont'Alverne C (2022) 'it's a battle you are never going to win': Perspectives from journalists in four countries on how digital media platforms undermine trust in news. Journalism Studies 23 pp 1—-20. URL https://doi.org/10.1080/1461670x.2022.2112908

US Defense D (2023) Strategy for operations in the information environment. URL https://media.defense.gov/2023/Nov/17/2003342901/-1/-1/1/2023-DEPARTMENT-OF-DEFENSE-STRATEGY-FOR-OPERATIONS-IN-THE-INFORMATION-ENVIRONMENT.PDF

Vosoughi S (2018) The spread of true and false news online. Science 359, no 6380 pp 1146—-51. URL https://doi.org/10.1126/science.aap9559

Wanless A (2025) The Information Animal: Humans, Technology and the Competition for Reality, Oxford University Press. Oxford University Press

Wardle C, Derakhshan H (2017) Information disorder : Toward an interdisciplinary framework for research and policy making. Council of Europe

Digital Curtains Descending: Measuring and Comparing Legal Closure over Digital Spheres after *TikTok v. Garland*

Davide Clementi

Department of Political Science and International Relations, University of Palermo, Italy.

Corresponding author. E-mail: davide.clementi@unipa.it

Abstract

This article examines how legal instruments and government actions following *TikTok v. Garland* have intensified forms of digital sovereignty, reshaping the boundaries of online spheres and increasing legal closure over cross-border digital infrastructures. The analysis compares national approaches to platform regulation, evaluates their compatibility with international legal norms, and explores their impact on digital freedoms and global internet governance.

Keywords

Digital Sovereignty, Platform Regulation, TikTok v. Garland, International Law, Digital Governance.

1 Introduction

The global digital landscape is undergoing a profound transformation, marked by increasing state intervention and a reassertion of national sovereignty in cyberspace. As Svantesson observes, 'while it has not always been so, today, it is uncontroversial to suggest that sovereignty applies online' (Svantesson, 2023), marking a clear departure away from the early ideals of a borderless and unregulated internet. This shift is driven by a confluence of factors including geopolitical rivalries, national security concerns, economic competition, and divergent ideological visions for the digital future. Already 2010, then-U.S. Secretary of State Hillary Rodham Clinton described 'virtual walls' that were 'cropping up in the place of visible walls', warning that government policies limiting the flow of information constituted 'a new information curtain [...] descending across much of the world'(Clinton, 2010). These 'digital walls' are digital barriers erected by governments that prevent or limit access to certain websites or online services, hindering the free flow of information across national borders (Vidyarthi and Hulvey, 2021).

A pivotal moment crystallizing this trend is the United States Supreme Court's 2025 decision in *TikTok v. Garland* (Supreme Court of the United States of America, 2025), a case poised to have far-reaching implications for how nations, particularly in the Western hemisphere, navigate the complex interplay of technology, governance, and fundamental rights. This paper analyzes the *TikTok v. Garland* decision as a harbinger of "Digital Curtains" of a new era, that of 'Digital Curtains'. Distinct from both impenetrable 'walls' and the clear logic of Internet 'borderlessness' (Svantesson, 2004)(Goldsmith, 2007)(Slane, 2008)(Maier, 2010), Digital Curtains are best understood as filtering instruments. They render the digital space inherently porous and traversable, until specific filters – underpinned by legal frameworks – are applied to at once conceal and reveal forms of political control, national discrimination, and other state-led objectives. This study then explores the comparative regulatory approaches(Baldwin, Cave, and Lodge, 2011) and architectures emerging in the United States, the European Union, and China in response to this new paradigm.

2 The Catalyst: *TikTok v. Garland* and the PAFACA

The *TikTok v. Garland* decision, which upheld the constitutionality of *the Protecting Americans from Foreign Adversary Controlled Applications Act* (PAFACA)(118th United States Congress, 2024), is not merely a legal judgment concerning a single application, but a significant marker in the evolving narrative of global digital governance. It underscores a decisive shift by a major liberal democratic power toward prioritizing national security imperatives in the regulation of foreign-owned digital platforms, thereby setting a potentially influential precedent.

The PAFACA, signed into law on April 24, 2024, mandates that social networking services deemed "foreign adversary controlled applications" be banned within 270 days unless divested, with a possible 90-day extension. The Act explicitly identifies ByteDance Ltd. and its subsidiary TikTok as falling within this definition, setting a compliance deadline of January 19, 2025. This legislation marked the culmination of years of congressional efforts to address national security concerns surrounding TikTok(Lemieux, 2025).

The primary justification for PAFACA centered on national security concerns, particularly the risk that the Chinese government could access American user data collected by TikTok or manipulate the platform's algorithm for influence operations and propaganda dissemination. U.S. lawmakers frequently cited China's 2017 National Intelligence Law - which could compel Chinese companies like ByteDance to cooperate with state intelligence agencies - as a key risk factor. These concerns included the storage of sensitive U.S. user data on servers located in foreign adversary countries, the monitoring of physical locations, and the creation of dossiers for purposes of blackmail or espionage.

It is important to note, however, that allegations of TikTok or ByteDance interfering with or manipulating American public opinion remain unsubstantiated or have even been explicitly refuted (Volz, 2020)(Kaplan and Sapiezynski, 2024), unlike the well-known Cambridge Analytica affair in the European context. The Act's provisions are designed to sever the control of foreign adversaries over such applications, thereby mitigating these perceived threats. Recent empirical studies, however, suggest that TikTok's algorithm does engage in rapid and pronounced amplification of content aligned with users' interests - typically within the first 200 videos viewed - potentially fostering topic-specific echo chambers and reducing exposure to diverse perspectives (Baumann, Arora, Rahwan, and Czaplicka, 2025). While these dynamics do not in themselves prove manipulation, they raise critical questions about the

socio-algorithmic feedback loops that such platforms may involuntarily trigger, with implications for informational pluralism and cognitive autonomy.

The normalization of national security as a primary driver of digital policy is clearly evident here: PAFACA exemplifies a broader trend in which security concerns are increasingly invoked to justify substantial interventions in the digital economy. This is not merely about a single application; it signals a paradigm shift in which the potential for data misuse and foreign influence through digital platforms is treated as an exigent threat that may outweigh traditional commitments to open markets or freedom of expression.

The legislative alacrity and the judicial deference suggest a broad political consensus: national security imperatives may justify robust - even preemptive - regulatory action against foreign-controlled applications, thereby setting a precedent for future measures targeting other platforms or technologies perceived as threats.

Following the enactment of PAFACA, TikTok and its parent company, ByteDance Ltd., mounted a legal challenge, arguing that the law violated several constitutional provisions. Their core arguments claimed that PAFACA infringed the Free Speech Clause of the First Amendment, constituted a prohibited bill of attainder under Article I, Section 9, and violated both the Due Process Clause and the Takings Clause of the Fifth Amendment. While Congress may enact legislation targeting corporations deemed threats to national security, it cannot pass laws that specifically punish an individual entity - so-called *Bills of attainder* (Wormuth, 1950)(Vidyarthi and Hulvey, 2021)(Berger, 1977)(Steilen, 2015). TikTok contended that a forced sale or outright ban constituted an unconstitutional restriction on its own speech and that of its millions of American users, asserting that the government's national security concerns were "hypothetical" and that less restrictive alternatives had not been adequately considered. It further argued that the law unfairly singled the company out for punitive treatment.

The U.S. government's defense, presented before the D.C. Circuit Court of Appeals(United States Court of Appeals for the District of Columbia Circuit, 2024), which unanimously rejected TikTok's claims in December 2024, emphasized that PAFACA was a national security regulation, not a speech regulation. The government argued that the law was content-neutral, targeting the conduct of foreign ownership and control by an adversary state, rather than the content disseminated on the platform. This framing aimed to subject the Act to intermediate scrutiny (Bhagwat, 2007)(Wexler, 1997)(Kelso, 2021) rather than the strict scrutiny usually applied to content-based speech restrictions. In support of its case, the government stressed the risks of covert data harvesting and algorithmic manipulation by a foreign adversary, submitting classified evidence to substantiate these claims.

In *TikTok v. Garland*, the Supreme Court upheld PAFACA. The Court held that - even assuming the Act implicated First Amendment rights - it withstood intermediate scrutiny(Supreme Court of the United States of America, 2025), given the compelling national security interests articulated by the government. The opinion of the Court endorsed the content-neutral rationale, focusing on the threat posed by a foreign adversary's control over a widely used application and its vast stores of user data, rather than the specific speech hosted on the platform. Some justices opined that strict scrutiny would have been appropriate, but concluded that the government had satisfied even that elevated standard, in light of the "vast troves of personal information" at stake. The ruling was likely narrowly tailored to the specific circumstances of the case and the imminent statutory compliance deadline.

The Supreme Court's validation of PAFACA - particularly under a deferential national security lens and an intermediate scrutiny framework - sets a significant precedent. It signals a judiciary increasingly willing to accord substantial weight to legislative and executive assessments of national security threats in the digital sphere, even where fundamental rights are at issue. The Court's acceptance of a content-neutral framing, focused on the nature of control (*i.e.*, foreign adversary ownership) rather than on expressive content, may provide a template for future regulatory interventions targeting foreign technologies.

3 Conceptualizing and Measuring Digital Curtains

3.1 A Fuzzy Logic Approach to Digital Sovereignty

The *TikTok v. Garland* decision and the broader trends it represents can be conceptualized through the metaphor of Digital Curtains: regulatory instruments enabling state intervention in cyberspace. These are not monolithic barriers but a spectrum of measures designed to assert national control over the

digital environment. This characterization as a spectrum immediately presents a significant methodological challenge for comparative analysis. A traditional approach, rooted in classical or Aristotelian logic, would attempt to frame the inquiry in binary terms: a Digital Curtain is either present or absent, a legal system is either open or closed. Such a framework is grounded in a certainty-oriented duality of "either this/or that", where propositions are strictly true or false. However, applying this binary logic to complex, nuanced phenomena like state control over the Internet is inadequate. It forces multifaceted realities into overly simplistic categories, abandoning the approximate arguments that characterize the real world because it lacks the tools to analyze them. To treat Digital Curtains as something that can be assessed with a simple yes/no answer would be an "offence to logic" and would fail to capture the very essence of these instruments as a graduated, multi-faceted assertion of sovereignty. To overcome the limitations of a binary framework, this study adopts a methodology explicitly grounded in fuzzy logic. Fuzzy logic consists in a form of many-valued logic designed to handle the concept of partial truth, where the truth value of a variable can be any real number between a set range (e.g., 0 and 1)(Mazzarese, 1997)(Menski, 2010). It is a logic of approximation that acknowledges that 'everything can be graded' and is thus more compatible with the complex and layered nature of social and legal phenomena(Schneider and Wagemann, 2010). The concept of Digital Curtains represents a form of what can be termed "institutionally-generated vagueness"(Perez, 2015), where the normative force, or "bindingness," of the regulatory regime is not absolute but can be realized in degrees. Therefore, instead of employing rigid "monothetic" categories, where a set of necessary and sufficient features must be present for inclusion, this analysis utilizes a "polythetic" or fuzzy approach(Perez, 2015). This allows us to conceive of legal systems as belonging to a category "to a certain extent", moving beyond the simple question of whether a Digital Curtain has descended to the more insightful question of to what degree. The empirical test outlined below is the operationalization of this fuzzy approach. It is designed to construct a Digital Curtain Index (DCI), which quantifies the degree of membership of a given jurisdiction within the category of having (or not) a descended Digital Curtain. This methodology mirrors the logic of fuzzy-set qualitative comparative analysis (fsQCA)(Kraus, Ribeiro-Soriano, and Schüssler, 2018), which systematically represents similarities and differences across cases to examine complex patterns. Rather than yielding a binary outcome, the test assesses a country's legal and regulatory framework against a set of defined criteria - Data Localization(Casalini and González, 2019)(Chander, 2020), Content Moderation, Algorithmic Oversight, and Economic Barriers - and assigns a calibrated score. This score represents the system's position on a continuum from a highly permeable digital environment (low DCI) to a comprehensively filtered one (high DCI). In doing so, the test provides a more powerful and accurate tool for analysis. It acknowledges that legal and regulatory instruments are not always crisp and absolute, but are often better understood as possessing varying shades of intensity and application. By embracing the principles of fuzzy logic, this framework moves beyond a simplistic and potentially misleading binary analysis, offering instead a structured, replicable, and nuanced method for comparing the diverse and evolving strategies of digital sovereignty across different legal systems.

3.2 The Digital Curtain Index (DCI)

Building upon the rationale for a fuzzy, degree-based analysis, this section outlines the empirical framework for calculating the Digital Curtain Index (DCI). The DCI is a composite, falsifiable index designed to provide a structured and replicable method for assessing the extent to which a state has implemented control mechanisms over its domestic digital environment. The final score is derived from the interaction between the intensity of the regulatory measures in place and the level of legal accountability governing them. The index is calculated using the following formula:

$$\text{DCI} = \text{Regulatory Intensity Score (RIS)} \times \text{Accountability Modifier (AM)}$$

The Regulatory Intensity Score (RIS) provides a quantitative measure of the control instruments, while the Accountability Modifier (AM) acts as a qualitative multiplier that adjusts this score based on the presence or absence of due process and transparency, reflecting the "porosity" of the curtain. The DCI is a positive, yet neutral index: it records the intensity and configuration of legal closure without implying normative endorsement or critique. The four RIS domains are not morally equivalent - some measures may pursue constitutionally legitimate purposes - yet the DCI deliberately abstracts from such evaluative distinctions. Its role is to capture how closure is structured and how far it descends, while its justifiability depends on the surrounding guarantees, thematised through the AM. It is important to acknowledge

that, despite the detailed articulation of the criteria, the process of calibrating the indicators for the DCI calculation is not entirely immune to an element of interpretive judgment. The complexity and nuances of real-world legal and regulatory frameworks imply that their subsumption into predefined ordinal categories necessarily involves approximation. To mitigate such inherent subjectivity, this study has relied on a rigorous and transparent adherence to the descriptions provided for each score level, for both the RIS and the AM. The objective is therefore not to provide an absolute measurement, but rather a replicable and systematic heuristic model, whose validity lies in the consistent application of explicit criteria and in its capacity to make comparable phenomena that would otherwise remain difficult to measure.

Indicator	Score	Description
A. Data Localization	0	No specific mandates; standard cross-border data flow rules apply.
	1	Sector-specific mandates (e.g., health, finance, gov. data).
	2	General localization with exceptions or transfer mechanisms.
	3	Cross-sectoral mandates; local storage and presence required.
B. Content Moderation	0	Safe harbor; only court-ordered takedowns for illegal content.
	1	Defined takedown obligations (e.g., NetzDG); predictable penalties.
	2	Vague categories (e.g., "harmful"); short deadlines; filtering.
	3	Direct state removal authority; harsh sanctions; service blocking.
C. Algorithmic Oversight	0	No specific oversight; general consumer/competition law only.
	1	Transparency and user opt-out from personalized feeds.
	2	Risk audits required; platforms must mitigate algorithmic harms.
	3	Direct state control over outcomes (e.g., ideological alignment).
D. Econ. & Tech. Barriers	0	Open market; level playing field for foreign and domestic firms.
	1	Indirect barriers (e.g., taxes, complex standards).
	2	Preferential treatment of national champions via subsidies etc.
	3	Bans or forced divestiture laws with severe penalties.

Table 1: Indicators for the Digital Curtain Index (DCI)

3.2.1 Component I: Regulatory Intensity Score (RIS)

The RIS measures the breadth and stringency of a state's digital control mechanisms across four key domains. Each domain is scored on an ordinal scale from 0 (minimal or no intervention) to 3 (pervasive and stringent intervention). The total RIS is the sum of these scores, with a maximum possible value of 12.

A. Data Localization Requirements (Score: 0–3)

This indicator assesses legal mandates that compel the storage of citizen or sectoral data within a country's territorial borders(Kuner, 2023).

0 (No Requirements): No specific mandates; cross-border data flows are governed by standard international agreements or corporate policies.

1 (Sector-Specific Requirements): Mandates are restricted to specific, sensitive sectors such as finance, public health, or government data.

2 (Broad Requirements): A general mandate exists requiring the localization of most categories of citizens' personal data, but with clear and accessible mechanisms for exceptions or transfers.

3 (Pervasive Mandates): A stringent, cross-sectoral legal requirement compels nearly all foreign and domestic companies to store citizen data locally and often to establish a local physical presence, mirroring the models in China or Vietnam.

B. Content Moderation Mandates (Score: 0–3)

This indicator measures the extent to which the state legally obligates internet intermediaries (social media platforms, ISPs) to police, filter, and remove content.

0 (Robust Safe Harbor): Intermediaries are broadly shielded from liability for user-generated content and are only required to act upon receipt of a valid judicial order for manifestly illegal content.

1 (Specific Takedown Obligations): The state mandates the removal of clearly defined categories of illegal content (e.g., terrorist propaganda, hate speech) within specific timeframes, as exemplified by Germany's NetzDG. Fines for non-compliance are significant but predictable.

2 (Broad Takedown Obligations): Mandates cover vaguely defined categories of content ("harmful," "anti-state," or "false information"). Takedown deadlines are extremely short, and platforms are pressured to implement proactive algorithmic filtering systems.

3 (Direct State Control): The state apparatus (e.g., a ministry or regulatory agency) has the direct, legally-backed authority to order the immediate removal of content with limited or no independent oversight, and failure to comply results in severe penalties, including service blocking.

C. Algorithmic Oversight Mechanisms (Score: 0–3)

This indicator assesses state efforts to regulate the design and output of the algorithms that curate and rank online information.

0 (No Specific Oversight): Algorithms are unregulated, subject only to general consumer protection or competition laws.

1 (Transparency and User Control): The law requires platforms to provide transparency about how their recommender systems work and to offer users meaningful opt-outs from personalized feeds.

2 (Process-Based Regulation): The state requires platforms to conduct and publish risk assessments and audits of their algorithms to mitigate societal harms (e.g., discrimination, radicalization) but does not dictate specific outcomes.

3 (Outcome-Based Regulation): The state directly regulates algorithmic outcomes, requiring providers to align their services with national objectives (e.g., promoting "socialist core values"), ensure content diversity, or obtain prior approval for certain algorithmic systems, as seen in China's regulations.

D. Economic and Technological Barriers (Score: 0–3)

This indicator measures the use of economic and regulatory tools to control the digital market, often favoring domestic entities over foreign ones.

0 (Open Market): Foreign and domestic technology firms compete on a level playing field with no significant, discriminatory regulatory barriers.

1 (Indirect Barriers): High regulatory compliance costs, digital services taxes, or complex standards disproportionately impact foreign entities, acting as de facto trade barriers.

2 (Preferential Treatment): The state actively promotes "national champions" through subsidies, biased public procurement, and other policies designed to disadvantage foreign competitors.

3 (Exclusionary Measures): The state employs outright bans on foreign applications or enacts laws enabling the forced divestiture of foreign-owned digital assets as a condition of market access, with prohibitive penalties for non-compliance.

3.2.2 Component II: Accountability Modifier (AM)

The RIS alone is insufficient, as the same level of control can be experienced very differently depending on the legal safeguards in place. The **Accountability Modifier (AM)** is a multiplier that adjusts the RIS to reflect the degree of legal accountability. It assesses the "procedural fairness" of the system, recognizing that a lack of due process and transparency makes a Digital Curtain heavier and more opaque. The AM ranges from **1.0** (high accountability) to **2.0** (no accountability):

1.0 (High Accountability): The relevant laws are clear, precise, and publicly accessible. Enforcement actions are subject to prompt and effective review by an independent judiciary. There are meaningful, accessible avenues of appeal for both individuals and corporations.

1.25 (Moderate Accountability): Laws are generally clear, but enforcement is primarily managed by executive or administrative bodies with judicial review being limited in scope or difficult to obtain.

1.5 (Low Accountability): The legal framework is characterized by vague or "extravagantly vague concepts", granting wide discretion to enforcement agencies. The process is opaque, and avenues

for appeal are limited or ineffective, creating a system that some might describe as a "due process wasteland".

1.75 (Very Low Accountability): Pervasive procedural opacity and established institutional biases that effectively restrict access to independent due process of law, while formal requirements for reasoning and some review channels remain.

2.0 (No Accountability): Decisions are made by political bodies without clear legal justification, public transparency, or any meaningful, independent appeals process. The system operates arbitrarily and is entirely controlled by political power, rendering the legal system a mere "façade of legality".

3.2.3. Interpretation of the Final DCI Score

The final **Digital Curtain Index (DCI)** score, calculated by multiplying the RIS by the AM, provides an ordinal measure of the "descent" of a country's Digital Curtain. The scores can be interpreted along a continuum:

- **Low DCI (e.g., 0–5): Slightly descended Curtain.** The digital environment is largely open. State interventions are targeted, limited, and governed by a system with high legal accountability.
- **Medium DCI (e.g., 6–12): Partially descended Curtain.** The state exercises noticeable filtering and control across several domains. While not a fully closed system, the effects of digital sovereignty are clearly felt by users and companies.
- **High DCI (e.g., 13–24): Fully Descended Curtain.** The state employs a comprehensive and often opaque suite of controls to create a heavily filtered, managed, and potentially "balkanized" digital ecosystem, where national objectives override principles of openness.

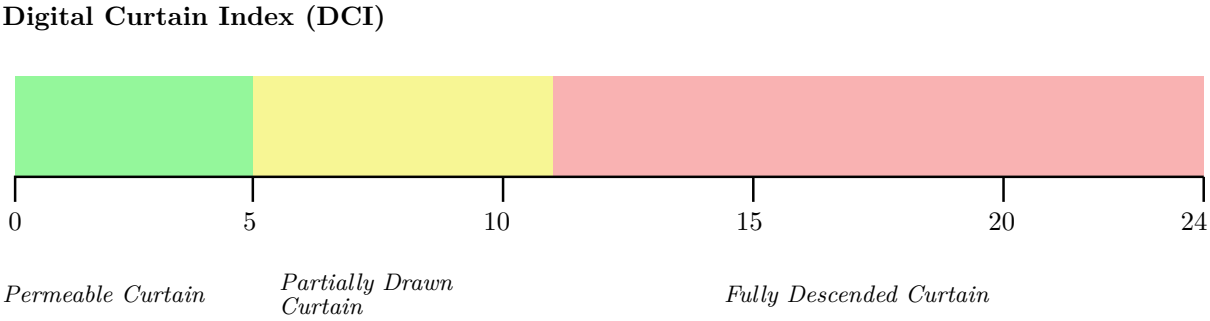


Fig. 1: Interpretation of DCI Scores along the Continuum

4 Calculating the DCI for the United States of America, the European Union, and the People's Republic of China

This section applies the DCI methodological framework to assess and compare the degree of "descent" of Digital Curtains in the United States, the European Union, and China.

4.1 United States of America

The U.S. approach is characterized by a growing emphasis on national security, which justifies significant interventions in the digital market, all while maintaining a formal structure of constitutional guarantees.

Regulatory Intensity Score (RIS) Calculation

A. Data Localization (2/3): The Department of Justice's Data Security Program, effective April 2025, establishes de facto export controls on Americans' sensitive personal data to "countries of concern"(U.S. Department of Justice, National Security Division, 2025). While not a generalized localization mandate, it represents a broad and stringent restriction on cross-border data flows for

national security reasons. The CLOUD Act also asserts U.S. extraterritorial jurisdiction over data held by American providers, regardless of its physical location(Rojszczak, 2020).

B. Content Moderation (2/3): The U.S. approach, embodied by PAFACA, is becoming extreme. Instead of regulating specific content, it imposes a total ban or forced divestiture of an entire platform, representing a drastic form of moderation, even if cloaked under the mantle of national security and framed by both Congress and the Executive as a disconnection between foreign adversary ownership and actual content moderation. This mechanism grants the state absolute control over an application's existence in the U.S. market if designed as ""foreign adversary controlled application". While this trajectory appears increasingly uncompromising, the strong emphasis on the First Amendment — particularly as reaffirmed in the concurring opinions of Justices Gorsuch and Sotomayor — may still represent a potential stumbling block to the emergence of a content governance model resembling that of China.

C. Algorithmic Oversight (2/3): PAFACA demonstrates a deep concern for algorithmic manipulation and influence. A "qualified divestiture" explicitly requires severing any operational relationship regarding content recommendation algorithms. This constitutes a direct and substantial state intervention in the algorithmic governance of a company, representing a strong process- and origin-based regulation.

D. Economic Barriers (3/3): This is the most explicit domain of U.S. intervention. PAFACA (ban/forced divestiture), the foreign investment screening role of CFIUS, and strategic export controls on advanced technologies like semiconductors constitute a formidable, proactive, and targeted set of barriers that define a "techno-nationalist" strategy(Capri, 2024)(Diegues and Roselino, 2023).

RIS Score (United States) $= 2 + 2 + 2 + 3 = 9$

Accountability Modifier (AM)

Assessment (1.25): The U.S. system exhibits a strong tension between national security and First Amendment rights. Although a judicial process exists, courts have (hypothetically) upheld PAFACA by applying a less demanding "intermediate scrutiny" standard after accepting the "content-neutral" argument. There is a tendency to accord significant deference to the executive branch's assessments of national security threats. This indicates a system with formal guarantees but practically limited judicial oversight. This corresponds to *Moderate Accountability*.

Interpretation and Final DCI

DCI Calculation (United States) $= 9$ (RIS) $\times$ 1.25 (AM) $=$ **11.25**

With a score of 11.25, the United States falls into the *Partially Drawn Curtain* category, but at its upper limit, nearly crossing into the next. This reflects a system that, while maintaining the forms of a state under the rule of law, has adopted extremely powerful and pervasive control instruments, justified by a broad and deferential interpretation of national security.

4.2 The European Union: A Rights-Based Regulatory Bastion

The EU's approach is defined by a horizontal and comprehensive legislative framework that aims to create a digital single market based on the protection of fundamental rights, fair competition, and democratic values, although it faces considerable enforcement challenges(Synodinou, Jougleux, Markou, and Prastitou, 2021)(Manganelli and Nicita, 2022)(Bertarini, 2023)(Neergaard and de Vries, 2024).

Regulatory Intensity Score (RIS) Calculation

A. Data Localization (1/3): The GDPR has a significant extraterritorial reach for data protection but does not impose a generalized localization mandate. The Gaia-X initiative aims to create a European data infrastructure to reduce dependency on non-EU providers and enhance "strategic autonomy", but it remains an initiative rather than a binding legal requirement(Catanzariti, 2022).

B. Content Moderation (2/3): The Digital Services Act (DSA) imposes comprehensive due diligence obligations, especially for Very Large Online Platforms (VLOPs), to tackle illegal content. It provides for clear "notice-and-action" procedures and appeal mechanisms, balancing control with freedom of expression(Leiser, 2023)(Kaushal, Vinhuy, Khan, Tyagi, Upadhyay, Basu, Reiman, De Filippi, and Ghosh, 2024).

C. Algorithmic Oversight (2/3): The AI Act adopts a risk-based framework with strict requirements for "high-risk" systems. The Digital Markets Act (DMA) imposes specific obligations on gatekeepers to ensure algorithmic fairness and prevent self-preferencing. This represents robust process-based regulation(Wachter, 2023)(Laux, Wachter, and Mittelstadt, 2024)(Martens, 2024).

D. Economic Barriers (1/3): The DMA is designed to ensure fairness and contestability, not to exclude foreign actors. Although it affects global tech giants, its primary goal is competition, not protectionism(Heidebrecht, 2024).

RIS Score (EU) $= 1 + 2 + 2 + 1 = 6$

Accountability Modifier (AM)

Assessment (1.0): The EU regulatory framework is explicitly grounded in the protection of fundamental rights guaranteed by the EU Charter. The Court of Justice of the European Union (CJEU) ensures that any limitation is necessary and proportionate. Despite some enforcement challenges, this system offers the highest level of legal clarity and judicial oversight. This corresponds to *High Accountability*.

Interpretation and Final DCI

DCI Calculation (EU) $= 6$ (RIS) $\times$ 1.0 (AM) $=$ **6.0**

With a score of 6.0, the European Union falls into the *Partially Drawn Curtain* category. Its approach is significantly regulatory but tempered by a robust framework of fundamental rights, distinguishing it clearly from the U.S. and Chinese models.

4.3 People's Republic of China

The Chinese model represents the most explicit and pervasive approach to digital governance, where technology is a fundamental tool for achieving state goals of social control, ideological conformity, and technological self-sufficiency(Creemers, 2022)(Creemers, 2023)(Creemers, Papagianneas, and Knight, 2024)(Clementi, 2024a).

Regulatory Intensity Score (RIS) Calculation

A. Data Localization (3/3): The 2017 Cybersecurity Law (CSL) introduced explicit data localization obligations, requiring data to be stored within China and subjecting cross-border transfers to security assessments(Hong, 2020)(Li, 2021)(Liu, 2022)(Peng, Shao, and Zheng, 2023).

B. Content Moderation (3/3): The Chinese system exercises strict censorship to ensure alignment with Communisty Party of China (CPC) ideology. Regulations explicitly promote "Socialist core values" (shèhuì zhǔyì héxīn jiàzhíguān), representing direct and total state control(Xu, Gong, and Yin, 2022)(Xu and Flew, 2022)(Clementi, 2024a).

C. Algorithmic Oversight (3/3): China enforces comprehensive algorithmic governance. The 2022 "Algorithmic Recommendation Management Provisions" and 2023 "Deep Synthesis" rules regulate outputs to serve CPC objectives, including social stability and ideological alignment(Xu and Flew, 2022)(Abiri and Huang, 2022)(Chen and Lunt, 2021).

D. Economic Barriers (3/3): Under policies like "Made in China 2025", China pursues technological self-sufficiency. The state directs investments and promotes national champions, creating a state-controlled digital ecosystem(Qian, 2025).

RIS Score (China) $= 3 + 3 + 3 + 3 = 12$

Accountability Modifier (AM)

Assessment (2.0): Although the Personal Information Protection Law (PIPL) shares some formal similarities with GDPR(Calzada, 2022)(Ong, 2023)(Clementi, 2024b), it operates in a legal environment without independent judicial oversight. Transparency is directed to state regulators, not to the public. The system serves Party priorities, not individual rights. This corresponds to *No Accountability*.

Interpretation and Final DCI

DCI Calculation (China) $= 12$ (RIS) $\times$ 2.0 (AM) $=$ **24.0**

With a maximum score of 24.0, China clearly falls within the *Fully Descended Curtain* category. Its system represents the most holistic implementation of a Digital Curtain, underpinned by pervasive state control and a legal framework designed to reinforce rather than limit it.

5 Conclusion

This paper has proposed a conceptual and empirical framework for assessing the intensity and legal architecture of state control over digital environments through the notion of Digital Curtains. By moving beyond binary classifications and adopting a fuzzy-logic-based metric, the DCI operationalizes the concept as a continuum, allowing for a differentiated analysis of national legal strategies and tendencies of digital sovereignty.

The DCI is explicitly designed to be falsifiable, that is, open to revision, refutation, and improvement. It provides a replicable and transparent method for comparative analysis, but it does not claim to exhaust the dimensions of governance that may become increasingly relevant in future evolutions of digital policy. As such, the model invites constructive critique and refinement. It can - and should - be extended to incorporate emerging regulatory domains such as quantum encryption, biometric surveillance, synthetic media governance, and decentralized digital infrastructures. The point is not to impose conceptual finality, but to build a heuristic that adapts to the evolving complexity of cyber-legal ecosystems.

Empirical validation across three paradigmatic jurisdictions – the United States, the European Union, and the People's Republic of China – has revealed a trend of strategic convergence. Each system exhibits distinct regulatory genealogies, ideological foundations, and constitutional orders. Yet, their current trajectories suggest a partial alignment around the imperative of reasserting state authority in cyberspace, though with vastly different degrees of transparency and rights protection.

In this respect, the DCI provides a lens to quantify and compare the varying intensities of closure within the distinct geopolitical spheres that has been described as "Three Digital Kingdoms"(Gao, 2023). While Gao highlights the divergence of legal-institutional regimes across these kingdoms, the logic that underpins their evolution increasingly points toward a shared structural tendency: the expansion of digital sovereignty through new regulatory enclosures that disrupt the existing legal order(Garibaldi and Deane, 2023)(Bradford, 2023). What emerges is not a cooperative divergence governed by trade agreements, but a unilateral reassertion of control—a closure of digital borders animated less by legal harmonization than by geopolitical antagonism. The fragmentation of cyberspace, therefore, appears to be driven not by a global regulatory consensus, but by a new form of digital mercantilism that builds barriers, severs flows, and undermines established commitments under international law. This shift is especially visible in the United States' techno-nationalist strategy, where economic de-globalization, strategic decoupling, and securitized digital policy have produced ripple effects beyond the Sino-American rivalry, impacting also the European Union and, by extension, global governance frameworks.

This is not the Brussels Effect (in which global firms internalize European norms)(Bradford, 2020), nor a Beijing Effect (the diffusion of authoritarian cyber-governance stemming from the socialist legality of the People's Republic)(Erie and Streinz, 2021), nor a revival of the Washington Consensus in digital form. Rather, what we are witnessing is the emergent logic of "adversarial convergence", a process driven less by normative imitation than by the pressures of economic de-globalization, strategic decoupling, and digital securitization.

These processes – accelerated by the techno-nationalist agenda of the United States under the Trump (and post-Trump) administrations – are reshaping the global digital order. In this landscape, the European Union, despite its normative commitment to fundamental rights, finds itself both targeted and compelled to recalibrate its regulatory model in response to extraterritorial constraints and geopolitical fragmentation.

As Krishnamurthy observes, the time may have come to question whether sovereignty – with its near-plenary authority over territorial domains – is still the most appropriate framework for understanding state power in the digital age(Krishnamurthy, 2025). The challenge is not merely to assert control, but to preserve the conditions for an open, free, interoperable, and secure Internet, without capitulating to the logic of enclosure.

The DCI, in this sense, does not celebrate closure. It diagnoses it. It does not reduce the complexity of digital governance to abstract models, but seeks to provide a pragmatic, scalable, and comparative instrument for understanding the evolving grammar of power in cyberspace. As Jackson warned(Jackson, 2019), while the law remains shrouded in the fog of innovation, the social and normative stakes of

disruptive technologies continue to grow. If digital curtains are descending, our task is not to tear them down blindly, but to understand their texture, thickness, and transparency, and to decide, as societies governed by law, to what extent we are willing to live behind them.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

118th United States Congress (2024) Protecting americans from foreign adversary controlled applications act. https://www.congress.gov/bill/118th-congress/house-bill/7521, public Law 118-50, 138 Stat. 955, enacted 24 April 2024; codified at 15 U.S.C. § 9901

Abiri G, Huang X (2022) The people's (republic) algorithms. Notre Dame Journal of International & Comparative Law 12:16–

Baldwin R, Cave M, Lodge M (2011) Understanding Regulation: Theory, Strategy, and Practice. Oxford University Press, Oxford

Baumann F, Arora N, Rahwan I, et al (2025) Dynamics of algorithmic content amplification on tiktok. arXiv:2503.20231 [cs.SI], URL https://arxiv.org/abs/2503.20231, preprint available at https://arxiv.org/abs/2503.20231

Berger R (1977) Bills of attainder: A study of amendment by the court. Cornell Law Review 63:355–407

Bertarini B (2023) European Union Digital Single Market: Legal Framework and Challenges. Collana di Diritto, FrancoAngeli, Milano, URL https://www.francoangeli.it/Libro/9788835152460

Bhagwat A (2007) The test that ate everything: Intermediate scrutiny in first amendment jurisprudence. University of Illinois Law Review pp 783–808

Bradford A (2020) The Brussels Effect: How the European Union Rules the World. Oxford University Press, Oxford

Bradford A (2023) Digital Empires: The Global Battle to Regulate Technology. Oxford University Press, Oxford, 10.1093/oso/9780197649268.001.0001

Calzada I (2022) Citizens' data privacy in china: The state of the art of the personal information protection law (pipl). Smart Cities 5(3):1129–1150. 10.3390/smartcities5030058

Capri A (2024) Techno-nationalism: How it's reshaping trade, geopolitics and society. John Wiley & Sons, Hoboken

Casalini F, González JL (2019) Trade and cross-border data flows. Tech. Rep. 220, Organisation for Economic Co-operation and Development (OECD), 10.1787/b2023a47-en, URL https://doi.org/10.1787/b2023a47-en

Catanzariti M (2022) Data sharing in search of sovereignty: The case of the european cloud. Network Industries Quarterly 24(3):16–20

Chander A (2020) Is data localization a solution for schrems ii? Journal of International Economic Law 23(3):771–784. 10.1093/jiel/jgaa024, URL https://doi.org/10.1093/jiel/jgaa024

Chen S, Lunt P (2021) Chinese Social Media: Face, Sociality and Civility. Digital Activism and Society: Politics, Economy and Culture in Network Communication, Emerald Publishing Limited, Bingley, 10.1108/978-1-83909-135-3

Clementi D (2024a) Between digital surveillance and individual protection: A juridical and comparative history of the cyberspace administration of china. Rivista di Digital Politics 4(2):343–370

Clementi D (2024b) La legge cinese sulla protezione delle informazioni personali. un gdpr con caratteristiche cinesi? Rivista di Diritti Comparati (1):189–216

Clinton HR (2010) Remarks on internet freedom, address before the newseum. URL https://2009-2017.state.gov/secretary/20092013clinton/rm/2010/01/135519.htm, uS Secretary of State, US Department of State, 21 January 2010

Creemers R (2022) China's cybersecurity regime: Securing the smart state. Available at SSRN, URL https://ssrn.com/abstract=4070682

Creemers R (2023) The chinese conception of cybersecurity: A conceptual, institutional, and regulatory genealogy. Journal of Contemporary China 33(146):173–188. 10.1080/10670564.2023.2196508, URL https://doi.org/10.1080/10670564.2023.2196508

Creemers R, Papagianneas S, Knight A (eds) (2024) The Emergence of China's Smart State. Rowman and Littlefield, Lanham, MD

Diegues AC, Roselino JE (2023) Industrial policy, techno-nationalism and industry 4.0: China-usa technology war. Brazilian Journal of Political Economy 43(1):5–25

Erie MS, Streinz T (2021) The beijing effect: China's digital silk road as transnational data governance. NYU Journal of International Law and Politics 54:1–92

Gao H (2023) Data sovereignty and trade agreements: Three digital kingdoms. In: Chander A, Sun H (eds) Data Sovereignty: From the Digital Silk Road to the Return of the State. Oxford University Press, New York, p 213–239, 10.1093/oso/9780197582794.003.0010

Garibaldi S, Deane F (2023) Cyberspace as a fifth dimension of national security: Trade measure exceptions. Journal of International Trade Law and Policy 22(2):67–88

Goldsmith J (2007) Who controls the internet? illusions of a borderless world. Strategic Direction 23(11)

Heidebrecht S (2024) From market liberalism to public intervention: Digital sovereignty and changing european union digital single market governance. JCMS: Journal of Common Market Studies 62(1):205–223. 10.1111/jcms.13458

Hong Y (2020) The institutional logic of security assessment of cross-border data transfers in china: Context and progress. International Cybersecurity Law Review 1(1–2):93–102. 10.1365/s43439-020-00007-2

Jackson BW (2019) Cybersecurity, privacy, and artificial intelligence: An examination of legal issues surrounding the european union general data protection regulation and autonomous network defense. Minnesota Journal of Law, Science & Technology 21:169–204

Kaplan L, Sapiezynski P (2024) Comprehensively auditing the tiktok mobile app. In: Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion). ACM, Singapore, Singapore, pp 1198–1201, 10.1145/3589335.3651260

Kaushal R, Vinhuy T, Khan S, et al (2024) Automated transparency: A legal and empirical analysis of the digital services act transparency database. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT). Association for Computing Machinery, New York, NY, USA, pp 1121–1132, 10.1145/3593013.3594077

Kelso RR (2021) The structure of intermediate review. Lewis & Clark Law Review 25:691–728

Kraus S, Ribeiro-Soriano D, Schüssler M (2018) Fuzzy-set qualitative comparative analysis (fsqca) in entrepreneurship and innovation research – the rise of a method. International Entrepreneurship and Management Journal 14:15–33. 10.1007/s11365-017-0461-8

Krishnamurthy V (2025) Anchoring digital sovereignty. Chicago Journal of International Law 25(2):417–477

Kuner C (2023) Data and extraterritoriality. In: Ryngaert C, Erp JN, Molenaar R (eds) Research Handbook on Extraterritoriality in International Law. Edward Elgar Publishing, Cheltenham, UK, p 356–371

Laux J, Wachter S, Mittelstadt B (2024) Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. Regulation & Governance 18(1):3–32. 10.1111/rego.12489

Leiser MR (2023) Analysing the european union's digital services act provisions for the curtailment of fake news, disinformation, & online manipulation. Disinformation & Online Manipulation April 24, 2023. Available at SSRN: https://ssrn.com/abstract=4420161

Lemieux P (2025) Tiktok, public choice, and the theater of the absurd. Regulation 48(1):3–5

Li Y (2021) La disciplina cinese del trasferimento transfrontaliero dei dati. Rivista Italiana di Informatica e Diritto 3(1):67–78. 10.32091/RIID0028

Liu J (2022) China's data localization. In: Routledge (ed) China's Globalizing Internet. Routledge, London, p 83–102

Maier B (2010) How has the law attempted to tackle the borderless nature of the internet? International Journal of Law and Information Technology 18(2):142–175

Manganelli A, Nicita A (2022) Digital fundamental rights in the eu. In: Bostoen F, de Streel A, Van Gorp N (eds) Regulating Digital Markets: The European Approach. Springer International Publishing, Cham, p 73–104, 10.1007/978-3-031-13674-4_4

Martens B (2024) The european union ai act: Premature or precocious regulation? Bruegel | The Brussels-based Economic Think Tank (8). https://www.bruegel.org/blog-post/european-union-ai-act-premature-or-precocious-regulation

Mazzarese T (1997) Fuzzy logic and judicial decision-making: The peril of a rationalist fallacy. In: Garzón Valdés E, et al (eds) Normative Systems in Legal and Moral Theory. Festschrift for Carlos E. Alchourrón and Eugenio Bulygin. Duncker und Humblot, Berlin, p 13–36

Menski W (2010) Fuzzy law and the boundaries of secularism. Potchefstroom Electronic Law Journal 13(3):1–18

Neergaard U, de Vries S (2024) The interaction between free movement law and fundamental rights in the (digital) internal market. In: Krajewski M, Malinauskaite J, Van de Gronden JW (eds) New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights. Springer Nature Switzerland, Cham, p 97–115, 10.1007/978-3-031-44647-8_6

Ong R (2023) Privacy and personal information protection in china's all-seeing state. International Journal of Law and Information Technology 31(4):349–375. 10.1093/ijlit/eaad011

Peng C, Shao G, Zheng W (2023) China's emerging legal regime for privacy and personal information protection. Tsinghua China Law Review 15. https://www.tsinghuachinalawreview.law.tsinghua.edu.cn/issues/info/10300

Perez O (2015) Fuzzy law: A theory of quasi-legal systems. Canadian Journal of Law & Jurisprudence 28(2):343–370

Qian J (2025) Governing China's Digital Transformation: Industrial Policy, Regulatory Governance, and Innovation. Routledge Studies in the Economics of Innovation, Routledge, Abingdon, Oxon; New York, NY

Rojszczak M (2020) CLOUD act agreements from an EU perspective. Computer Law & Security Review 38:105442. 10.1016/j.clsr.2020.105442

Schneider CQ, Wagemann C (2010) Standards of good practice in qualitative comparative analysis (qca) and fuzzy sets. Comparative Sociology 9(3):397–418. 10.1163/156913210X12493538729793

Slane A (2008) Tales, techs, and territories: Private international law, globalization, and the legal construction of borderlessness on the internet. Law and Contemporary Problems 71(3):129–151

Steilen M (2015) Bills of attainder. Houston Law Review 53:767–832

Supreme Court of the United States of America (2025) Tiktok inc. v. garland. URL https://www.supremecourt.gov/opinions/24pdf/24-101_6k47.pdf, no. 24-101, decision delivered on 30 June 2025, reported at 600 U.S. — (2025)

Svantesson D (2004) Geo-location technologies and other means of placing borders on the 'borderless' internet. John Marshall Journal of Computer and Information Law 23:101–138

Svantesson D (2023) A starting point for re-thinking "sovereignty" for the online environment. In: N.N. (ed) Data Sovereignty: From the Digital Silk Road to the Return of the State. p 49–50

Synodinou TE, Jougleux P, Markou C, et al (eds) (2021) EU Internet Law in the Digital Single Market. 315690, Springer International Publishing, Cham

United States Court of Appeals for the District of Columbia Circuit (2024) Tiktok inc. and bytedance ltd. v. merrick b. garland. Argued 16 September 2024, decided 6 December 2024

U.S. Department of Justice, National Security Division (2025) Nsd data security program – compliance guide. URL https://www.justice.gov/nsd, version 2, April 11, 2025

Vidyarthi A, Hulvey R (2021) Building digital walls and making speech and internet freedom (or chinese technology) pay for it: An assessment of the us government's attempts to ban tiktok, wechat, and other chinese technology. Indian Journal of Law and Technology 17(1):1–42

Volz D (2020) U.s. national security adviser says china targeting 2020 election. URL https://www.wsj.com/articles/u-s-national-security-adviser-says-china-targeting-2020-election-11597007831, the Wall Street Journal, 9 August 2020

Wachter S (2023) Limitations and loopholes in the eu ai act and ai liability directives: What this means for the european union, the united states, and beyond. Yale Journal of Law & Technology 26:671–732

Wexler JD (1997) Defending the middle way: Intermediate scrutiny as judicial minimalism. George Washington Law Review 66:298–325

Wormuth FD (1950) Legislative disqualifications as bills of attainder. Vanderbilt Law Review 4:603–624

Xu J, Flew T (2022) Governing the algorithmic distribution of news in china: The case of jinri toutiao. In: Flew T, Martin F (eds) The Algorithmic Distribution of News: Policy Responses. Springer International Publishing, Cham, p 27–47, 10.1007/978-3-030-99926-5_2

Xu J, Gong Q, Yin W (2022) Maintaining ideological security and legitimacy in digital china: Governance of cyber historical nihilism. Media International Australia 185(1):26–40. 10.1177/1329878X221096163

The Demand for Bullshit Under American and European Law

Lila Greenberg[1], Julie Wittenberg[1], Jane Bambauer[2]

[1] University of Florida Levin College of Law, J.D. expected 2026.
[2] Brechner Eminent Scholar, University of Florida Levin College of Law and College of Journalism and Communications.

Corresponding authors. E-mails: lilagreenberg@ufl.edu; janebambauer@ufl.edu

Abstract

Our information environment is saturated with bullshit—not just fake news and disinformation, but also low-quality, ideologically slanted content that encourages lazy thinking. While critics often blame content creators or platforms, the real problem lies on the demand side: audiences seek out confirmatory content, pressuring media to cater to their biases. Current legal interventions, whether European-style regulation or U.S. attempts to limit youth exposure, miss this point and risk backfiring. This article argues that supply-side interventions will be ineffective or counterproductive, potentially increasing distrust and extremism. Instead, we propose two demand-side approaches: (1) treating media as a vice product requiring regulation, and (2) fostering critical thinking skills to reduce susceptibility to demagoguery. Both face implementation challenges—Americans need better education in logical reasoning, while Europeans may focus too heavily on their existing supply-side regulations—but offer more promising paths forward than current policy approaches.

Keywords

Information Environment, Bullshit, Law, Policy, Critical Thinking.

1 Introduction

Our media ecosystem is awash in bullshit. "Bullshit," as used in this paper, refers not only to outright disinformation or fake news, but to a broader category of misleading or low-quality content. This includes sloppy reporting, ideologically slanted narratives, and emotionally manipulative content that encourages lazy, confirmatory thinking rather than critical engagement. Unlike deliberate lies, bullshit is often unconcerned with truth altogether—it prioritizes identity affirmation or attention capture over accuracy. Drawing from Harry Frankfurt's famous account, bullshit bypasses the norms of honest discourse by disregarding whether what is said is true or false, focusing instead on what will satisfy an audience's desires, fears, or affiliations (Frankfurt, 2005). In the digital media ecosystem, where attention is scarce and algorithms reward engagement, bullshit thrives not because it deceives, but because it gratifies. In this article, "media" refers broadly to content on matters of public concern, whether produced by traditional news outlets, new media organizations, or individual social media users.

A dominant explanation for the spread of misinformation and low-quality content focuses on supply-side actors, particularly the role of new media companies and technology platforms. According to this view, social media algorithms amplify sensationalism, new media outlets chase clicks at the expense of accuracy, and digital platforms are designed to exploit their customers. These concerns have shaped the policy agenda in both the U.S. and the E.U. But they miss a central feature of the problem: we are awash in bullshit because we *like* it.

American and European lawmakers have tried two contrasting approaches to handle this problem. In the European Union, regulators have enforced laws against hate speech, and content-based regulatory schemes like the Digital Services Act require platforms to proactively monitor and remove unlawful content and explain their content moderation decisions. The U.S. government is constrained from doing either of these things under the strong libertarian interpretation of the First Amendment. However, many states are experimenting with laws that protect young users from harmful or addictive news feeds on social media, and the U.S. Congress has successfully (sort of) forced the sale of TikTok.

These legal interventions fail to fully recognize that the success of bullshit stems more from consumer demand than from producer or distributor supply. Social media home feeds, new media content creators, and even legacy media are under significant market pressure to cater to the organic, pre-existing social and political beliefs of their audiences.

This article gives audience capture and the demand side of the bullshit problem the serious legal treatment that it deserves. The legal interventions and proposals that are popular today, while intuitively appealing, will be ineffective or counterproductive. Unless consumer preferences change, legal solutions that interfere with supply are unlikely to work, and may even backfire, causing distrust, conspiratorial thinking, and political extremism.

We offer a sketch of two alternative approaches that the government can take to reduce the appeal of bullshit: (1) policies that treat media as a vice product (while remaining agnostic on *which particular claims* are dangerous); and (2) policies that foster a culture of critical thinking and distrust of demagoguery.

Both of these approaches will encounter tough sledding in the U.S. and the EU, as they are unlikely to garner political support or popular interest. However, the frictions and implementation challenges will be slightly different. American citizens start at a deficit in critical thinking and logical reasoning skills compared to their European counterparts, and large-scale changes in the education system are difficult in the decentralized school system in the United States. However, Americans may be more inclined to use market-oriented tools that penalize bullshit. And the EU may be distracted and delayed from trying promising demand-side interventions while they attempt to make their supply-side interventions work.

2 Why Are We Covered In Bullshit?

2.1 Supply-Side Theories

Supply side theories dominate legal and policy understanding of, and responses to, the bullshit problem.

2.1.1 Manipulation

A dominant theory in law and mass communications explains "bullshit" as the product of manipulation by media platforms. The theory assumes that platforms and content creators push false and misleading content to shape consumers' beliefs, emotions, and behaviors without their conscious awareness (Susser,

Roessler, and Nissenbaum, 2019). This "covert persuasion" model holds that media companies design platforms to maximize emotional reactions and tribal reinforcement, regardless of factual accuracy; such manipulation is intentional, exploitative, and harmful, treating attention as a commodity to be extracted for profit (Norton, 2021). However, when lawmakers ban TikTok or public intellectuals criticize podcast hosts for "platforming" bullshit, they base these criticisms on the assumption that bullshit spreads not because users want it, but because content creators and social media platforms are injecting unfavorable viewpoints into users' minds and conditioning users to want what the algorithm shows.

2.1.2 Filter Bubbles and Echo Chambers

Filter bubbles and echo chambers are central to the dominant understanding how social media enables the spread of misinformation and confirmatory "bullshit." User manipulation functions through siloed ideological groups, called echo chambers and filter bubbles. Users are algorithmically nudged into ideologically homogenous spaces that reinforce their existing beliefs and minimize exposure to dissenting views. Users' natural preference for confirmatory content is amplified by personalization and engagement-driven platform design, creating a self-reinforcing feedback loop that deepens ideological commitments. Within these closed systems, even contradictory evidence can fuel tribal identification, as opposing viewpoints are not just ignored but actively discredited (Nguyen, 2020). Emotional clarity and identity affirmation often outweigh factual accuracy, making misleading but affirming content more appealing and viral. Algorithms capitalize on our preference for validation and connection, subtly driving radicalization through emotional reinforcement (Langvardt, 2019). In this way, social media platforms don't merely host angry tribal politics, they intensify and reward them.

2.1.3 Failure to Screen

Many critics on both sides of the Atlantic argue that social media companies have failed in their duty to screen and moderate harmful content, allowing bullshit to flourish unchecked. This failure is not simply a technical oversight but instead reflects platforms' reluctance to comply with their ethical and functional responsibilities to users. Legal scholar Jack Balkin proposed that social media platforms act as "information fiduciaries," obligated to act in users' best interests by avoiding unexpected or exploitative behaviors, such as enabling bullshit to go viral. Others contend that platforms should serve as responsible "custodians of the internet," ensuring that platform design choices and policies do not facilitate the spread of bullshit (Gillespie, 2018). The absence of robust fact-checking and editorial filters is a key factor in disinformation's virality. Studies show that moderation verification practices can reduce the spread and influence of disinformation (Allcott and Gentzkow, 2017). When platforms underperform or retreat from content moderation, as seen with Meta's decision to scale back enforcement, they violate users' reasonable expectations that they can trust platforms to screen content for accuracy.

2.1.4 Bias in Screening

Supply-side theories have been used to push in the other direction as well. Conservatives in both the United States and the European Union have long expressed concerns that major technology platforms exhibit systemic bias against right-leaning viewpoints. In the U.S., these concerns intensified during the 2016 and 2020 election cycles, as prominent Republican lawmakers and commentators alleged that platforms such as Facebook, Twitter, and YouTube suppressed conservative speech through algorithmic demotion, content removals, and bans of high-profile figures. Similarly, conservative parties and politicians in the EU have raised alarms about the perceived ideological leanings of Silicon Valley–based companies and their influence over public discourse.

Thus, no matter where along the ideological range the criticism comes from, platform algorithms are blamed for injecting misleading or polarizing content into users' minds. But this framing misrepresents the underlying dynamic. In most cases, social media platforms are not pushing unwanted content onto unsuspecting audiences, and do not have much power to change beliefs. Instead, they are responding to user demand. The algorithms serve what users believe, and the platforms would face grave financial consequences if they failed to supply users with the content they desire.

2.2 Demand-Side Realities

A growing body of research undermines the common assumption that social media platforms are to blame for spreading bullshit. For the most part, bullshit seems to succeed when users want to consume it, and it doesn't when users don't.[3]

Studies demonstrate that the structure of content feeds, whether algorithmically curated or shown in reverse chronological order, has minimal impacts on political polarization (Guess et al., 2023). Additionally, users' exposure to false or inflammatory content is limited and highly concentrated among a small group of motivated users who actively seek it out (Budak et al., 2024; Hosseinmardi et al., 2021). The virality of disinformation is not just a symptom of manipulation, it is a reflection of user preferences. But why do users want bullshit? Given the importance of having a factually accurate account about how the world works, why would confirmation be so highly valued?

2.2.1 Social and Psychological Goals

Individuals are often motivated by goals unrelated to knowledge, often seeking social identity more than the truth, especially when their beliefs are central to their sense of self. Communities around shared ideologies naturally gravitate toward information, true or false, that reinforces those beliefs, with partisan affiliation serving as a powerful lens through which facts are filtered. There are strong practical reasons and psychological impulses that make it easier to accept misinformation that flatters one's ingroup and to reject accurate information that threatens it. Groups thrive with the help of content that reinforces loyalty rather than accuracy. This creates a socially stable but epistemically dysfunctional media ecosystem.

2.2.2 Epistemic Goals, Faulty Tools

Even the most motivated and epistemically conscientious news consumers face serious challenges in distinguishing truth from bullshit in today's saturated media environment. With content competing for attention across all platforms, individuals are required to evaluate conflicting claims rapidly and constantly. While this may lead to healthy skepticism, it also places an enormous cognitive burden on users—one that even the most analytically minded cannot always meet. According to dual process theory, our cognitive evaluations follow two tracks: intuitive (System 1) and analytical (System 2) reasoning (Kahneman, 2011). Most of the time, especially under the pressure of information overload, individuals default to System 1 processing, using heuristics and gut responses rather than rigorous scrutiny. This opens the door to accepting false or misleading claims, particularly when they appear familiar or emotionally resonant.

A segment of scholars and the general public argue that equipping media consumers with critical thinking skills will increase the overall desire for accurate content and dissuade consumers from bullshit. However, not all System 2 processing safeguards against bullshit. System 2 theories are divided into two accounts: classical reasoning, where analytical reasoning is applied in a neutral manner to make informed and accurate conclusions; and motivated reasoning, where analytical tools are used to defend identity or ideology. Thus, even the most analytical individuals may fall prey to bullshit if they apply their reasoning selectively to protect preexisting beliefs.

***

Thus, flawed reasoning and the prioritization of other goals make the demand for bullshit hard to tackle.

3 Supply-Side Legal Proposals

The following discussion highlights a representative set U.S. legal proposals alongside key enacted laws in the E.U., all of which have drawn sustained scholarly or legislative attention.

3.1 American Legal Proposals

Although the First Amendment places significant limits on regulating speech, U.S. lawmakers and scholars have explored various strategies to reduce the spread and impact of bullshit. These efforts underscore the contradictions of the American speech regime: since direct content regulation is largely off-limits,

[3] Empirical findings in this area face replication challenges. Thus, while we find the demand-side story convincing both as a matter of theory and empirical evidence, it may be more accurate to say the literature offers important but inconclusive evidence that user demand drives outcomes.

policymakers target the platforms and ecosystem where bullshit circulates. At both federal and state levels, proposals have been introduced or enacted to:

Increase Competition Among Platforms. Some advocates argue that increasing market competition would incentivize platforms to adopt better content moderation practices (Adams and Bambauer, 2026). The current dominance of a few large social media companies has limited the diversity of approaches to addressing misinformation. By reducing this dominance and fostering a more competitive ecosystem, smaller platforms can emerge and differentiate themselves by offering more trustworthy information environments. These platforms may compete more effectively by offering stronger commitments to accuracy and transparency in order to attract and retain users who value accurate information.

Promote Middleware. Middleware solutions aim to shift control over content moderation decisions from platforms to users. Middleware in the form of third-party software intermediaries between platforms and users allows users to customize their content feeds and moderation settings independently of the platforms themselves. This approach will decentralize control over the information environment and give users more autonomy to the opaque and often controversial decisions made by platforms (Fukuyama and Katz, 2021; Grimmelmann, 2023; Rozenshtein, 2021).

Revise or Remove Section 230 Immunity. Another widely debated proposal calls for revising or repealing Section 230 of the Communications Decency Act, a key provision that shields internet platforms from liability for user-generated content (Armijo, 2021). Modifying or eliminating this immunity would expose platforms to greater legal risk, potentially encouraging platforms to moderate harmful, defamatory or misleading content more aggressively.

Defective Design Claims / Social Media Age Limits. Lawmakers and scholars have proposed regulating media companies indirectly by relying on product liability theories, particularly claims based on design defects. These efforts treat features like infinite scroll and algorithmic recommendations as intentionally manipulative design choices that jeopardize users' psychological and physical health and erode democratic self-governance. Legal approaches under this theory aim to hold platforms accountable for deliberately addictive or dangerous design features, particularly those that harm younger users, thereby encouraging safer and more ethically responsible platform design (Stray et al., 2023; US, 2023b).

Outlaw "Surveillance Capitalism." Another proposal targets the underlying business model of "surveillance capitalism," in which social media platforms collect, analyze, and monetize vast amounts of personal data to serve targeted advertisements and maximize user engagement (New America, 2023). This model inherently incentivizes platforms to prioritize emotionally charged and polarizing content over epistemic quality. Legal reforms aimed at limiting data harvesting, restricting behavioral advertising, or requiring platforms to adopt alternative revenue models would reduce platforms' dependence on keeping users engaged at any cost, regardless of the quality of the content they host and promote.

Compelled Neutrality. Another popular legal proposal, although largely foreclosed by the U.S. Supreme Court's decision in *Moody v. NetChoice* (US, 2024), would require platforms to expose users to diverse viewpoints by imposing a legal duty to host and amplify content that might otherwise be demoted or removed (Volokh, 2021). These measures aim to restrict viewpoint-based moderation and reduce the formation of ideological echo chambers by forcing exposure to countering perspectives.

Mandatory Disclosures. Transparency-based proposals would require platforms to disclose key information about how their algorithms operate, how content moderation decisions are made, and the standards that guide enforcement actions. By enabling lawmakers and the public to evaluate how platforms handle misinformation, these measures aim to promote fairer, more consistent moderation while creating stronger incentives for responsible misinformation management (Florida, 2021).

Subsidized or Official Fact-Checking. Government-supported fact-checking initiatives would aim to reduce the spread and impact of false or misleading information (Bloch-Wehba, 2026; Langvardt, 2019). These efforts may include government-initiated "pre-bunking," which involves proactively addressing and debunking false claims before they gain traction. They may also involve providing authoritative corrections and verified information after misinformation has already circulated.

All of these proposals regulate what would be considered pure speech under American First Amendment law in some way—whether by compelling or obstructing it—so there are some limits on American proposals even beyond the political and practical challenges of passing legislation.

3.2 European Legal Proposals

Compared to the United States, the European Union has taken a more centralized regulatory approach to confront the disinformation problem. The differing cross-Atlantic approaches stem from the divergence

in speech norms in the U.S. and the EU. Whereas in the U.S., the First Amendment broadly protects speech, European speech regulations reflect the region's greater concern for speech-related harms, such as hate speech or reputational damage.

The EU has implemented a comprehensive framework of digital regulations, beginning with the e-Commerce Directive (EU, 2000) and the General Data Protection Regulation (EU, 2016), and more recently the Digital Markets Act (EU, 2022a) and Digital Services Act (EU, 2022b). In addition to these regulations, European policymakers have pursued other measures, but we focus on the DMA and DSA because they remain the most developed and enforceable frameworks with broad applicability across the digital ecosystem. In particular, the DSA represents a major step forward in establishing enforceable responsibilities for online platforms and creating a safer digital environment. The goal is to prevent illegal and harmful activities, including the spread of disinformation.

The DSA applies to online intermediary services and adopts a tiered approach, imposing heightened obligations on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). These obligations include conducting systemic risk assessments, implementing reasonable mitigation measures, and enhancing transparency, especially around recommender algorithms and content moderation practices.

Under the DSA, providers of intermediary services are generally exempt from liability for content hosted on their platforms as long as they lack actual knowledge of illegality or act swiftly to remove it once aware. Platforms must implement notice-and-takedown procedures that allow users to flag illegal content, while also providing clear explanations for any content removals or restrictions and informing users of appeal mechanisms. Additionally, intermediary services must moderate content accordance with clear and accessible terms and conditions, publish transparency reports on enforcement practices, and uphold procedural safeguards to ensure fair and consistent application of the terms.

The sister regulation to the DSA, the Digital Markets Act, aims to limit the dominance of large tech "gatekeepers" and foster fairer competition in digital markets. The DMA imposes "do's" and "don'ts" on gatekeepers. The obligations include, but are not limited to, ensuring inter-operability between gatekeepers and third-party products and services and providing business users access to data gathered or generated on the gatekeeper's platform. Gatekeepers are prevented from favoring their own products and services above other providers in search results or advertisements on the gatekeeper's platform. By imposing obligations and prohibitions on powerful platforms that control core online services, the Act is designed to open the market to smaller competitors.

Together, the DMA and DSA mark a decisive shift toward a more accountable and transparent digital environment. This twin regulatory framework signals the EU's effort to shape the digital space in ways that could limit the spread of disinformation. The EU has set high expectations, but the forces driving disinformation may prove more resilient than regulation alone.

3.3 The Limits of Supply-Side Legal Interventions

We suspect the supply-oriented legal interventions will have limited success. First, some of the proposals are designed to increase competition and provide consumers with more media options. These will exacerbate the problem by giving users even more of the bullshit they want.

The other interventions, which attempt to reduce the supply of fake news and confirmatory bullshit, will not only cause distrust and charges of bias but will also be ineffective. Unless governments are prepared to go to extraordinary lengths to constrain bullshit everywhere it may emerge, supply-side content suppression is unlikely to have a significant impact on beliefs or polarization. Users' own preferences and behaviors, rather than magnetic influencers or algorithmic manipulation, are the dominant force shaping their information environment.

4 Demand-Side Legal Responses

Can law and policy do more to address the demand for bullshit? Can law help reduce the social and psychological payoff for confirmatory bullshit, and increase an interest in truth-seeking?

4.1 Regulating Bullshit as a Universal Human Vice

One way to understand the problem of bullshit demand is to treat it as a vice good—that is, as a product that people consume even though they would prefer not too, *i.e.*, addiction, or even though it causes net welfare declines to the consumer and to others (Wertenbroch, 1998).

Some legal scholars have suggested regulating online misinformation as a vice that can be controlled through paternalistic restrictions (US, 2023a). We believe treating media consumption as a vice has potential, but current proposals miss the mark. First, the proposals often target only social media platforms or algorithms, and these interventions will fail if consumers can easily find bullshit elsewhere.

More importantly, so far the vice-based proposals have treated misinformation as an aberrational-but-severe affliction, like a cocaine addiction. This is descriptively wrong and a political non-starter. The demand for bullshit is a normal human foible that everyone shares to some degree, not an aberrant behavior requiring crisis-level intervention. If interventions are used that require some final authority on "truth," or an official account of which falsehoods are "dangerous," the political groups who hold beliefs that are targeted will become more hostile and suspicious of democratic institutions rather than less.

The key is approaching bullshit as a broadly-shared vice, more like junk food than cocaine. This means using interventions similar to those for other common vices—age limits, Pigouvian taxes, time and place restrictions, or perhaps reviving fairness doctrine requirements—rather than content-based restrictions.

4.2 Enabling the Private Solutions

The government can encourage a culture of truth-seeking and open-minded inquiry by supporting certain types of content that taps into audience truth-seeking or by supporting institutions that make bullshit and lazy thinking more costly—financially and socially.

4.2.1 Truth-Seeking Content Hacks

Content hacks are nudges designed to tap into a news consumer's own desires to be truth-seeking, or to dampen the instinct to be confirmation-seeking. These are often simple enough to be implemented at scale by social media platforms or news producers.

For example, quick little logic games like sudoku or riddles can put the audience in a problem-solving mindset rather than a battle mindset while they peruse the news (Salvi, Barr, Dunsmoor, and Grafman, 2023). Quiz questions asking readers about their background knowledge on a subject can help illuminate to the reader that they do not (or should not) have overwhelming confidence in the beliefs that they currently hold. And we are in the process of testing whether simple nudge messages such as "the world is complicated" can help increase the amount of time that a news consumer spends in a more exploratory, open-minded mode of reasoning.

Audience collaboration systems can harness collective intelligence to provide real-time corrections and additional context, and they can also produce content tagging that is less partisan and more broadly trusted. Community Notes exemplifies this approach by crowdsourcing fact-checking directly within social media platforms, allowing users to collaboratively annotate potentially misleading content with contextual information, but this particular system does not seem to be able to keep up with the volume and velocity of online bullshit (Chuai et al., 2024).

Cross-cutting exposure mechanisms deliberately present users with diverse viewpoints that challenge their existing beliefs, potentially reducing the echo chamber effects that fuel demand for confirmatory bullshit.

None of these will be silver bullets, but they may have sustained and cumulative effects that help reduce the love of bullshit.

4.2.2 Bullshit Penalties: Prediction Markets, Wagers, Private Guarantees

Wagers and prediction markets create financial incentives that can counteract the demand for bullshit by converting abstract political beliefs into concrete monetary decisions with direct personal stakes. While confirmation bias tends to be pronounced for identity-driven political beliefs, it diminishes when individuals face immediate financial consequences, directing attention toward better evidence and more realistic assessments. Prediction markets like Polymarket exemplify this principle by allowing politically charged events to be evaluated through financial markets that reward accuracy.

Truth bounties, as proposed by Yonathan Arbel and Michael Gilbert, offer another market-based mechanism where news producers can voluntarily guarantee their work by offering substantial payments to anyone who can prove important story elements false to a neutral arbitrator (Arbel and Gilbert, 2024; Lin, 2024). The government can promote or subsidize these innovations and ensure that legal obstacles don't impede their development.

4.3 Invest in Critical Thinking Education

The most straightforward way to induce demand for better-quality news is to foster a culture of critical thinking. The government usually does this, if it does it at all, through its influence over K-12 education. Government interventions could reform formal schooling by encouraging more engagement with complexity and requiring foundational knowledge in statistics, probability, causal reasoning, and informal logic—skills often more applicable to real-world decision-making than traditional requirements like algebra or trigonometry. Given the decentralized American K-12 system, interventions must be adaptable. Moreover, because adult reasoning skills can deteriorate over time, durable reforms should also include lifelong learning initiatives that reinforce epistemic habits beyond school. A growing body of research demonstrates that media literacy training produces measurable effects, with studies showing that "active inoculation" makes participants "more attuned to specific deception strategies" rather than merely more skeptical (Roozenbeek and van der Linden, 2019), and that information literacy significantly correlates with fake news identification (Jones-Jang et al., 2019).

However, research on system 2 human cognitive processes reveal a crucial limitation: training alone is insufficient, as partisans with enhanced statistical and logical skills often use these abilities to reinforce their existing biases and dig deeper into confirmation bias. Therefore, educational interventions must be coupled with broader cultural changes that celebrate modesty, precision, and accuracy on verifiable claims while fostering awareness of what leads to tribalism and demagoguery, including us-versus-them thinking and encouraging resentment, and actively ridiculing excessive partisanship to create social incentives for intellectual humility.

5 Differential Strengths in the US and EU

While demand-side interventions offer promising alternatives to supply-side content regulation, their effectiveness depends heavily on cultural and institutional contexts that vary significantly between the United States and European Union. The success of approaches like critical thinking education, decentralized content verification programs, and market-based truth verification mechanisms will be shaped by deep-seated differences in social trust, educational structures, and fundamental values. Understanding these cross-Atlantic divergences is essential for designing demand-side policies that can realistically succeed within each region's distinct political and cultural environment.

The Inglehart-Welzel study of world cultural values reveals significant differences between the United States and European countries that may affect their respective approaches to combating bullshit. European nations, particularly those in central and northern continental Europe, tend to have relatively high levels of trust and rationalism among their populations, suggesting they may have cultural foundations more conducive to evidence-based thinking and justified trust in experts. Factors such as relative economic equality, educational quality, social cohesion, and religious heritage may contribute to resilience against bullshit. In contrast, the United States routinely rates lower on both dimensions, indicating relatively lower levels of social trust and less emphasis on rationalist approaches to problem-solving.

This cultural difference suggests that Europeans may be more receptive to institutional solutions and expert-driven interventions, while Americans may be more skeptical of authority-based approaches to addressing misinformation. These cultural differences have important implications for policy design, as demand-side interventions that rely on trust in institutions (for example, changes to the education curriculum) may face greater resistance in the American context compared to European settings where such values are more deeply embedded in the social fabric. Europeans may also be starting with a slightly higher baseline in rational reasoning skills.

However, Americans likely have an edge in accepting new, weird, market-driven institutions like prediction markets and wager systems precisely because they are less attached to existing forms of expertise.

Differences between the educational systems in the EU and the U.S. also create distinct challenges for each region. Policymakers might want to reduce the demand for bullshit by implementing educational policies that strengthen epistemic practices and critical reasoning. In the U.S., the K-12 educational system is highly decentralized, with curricula and standards set at the state level. Despite this fragmentation, school systems across the U.S. generally rely on multiple-choice and recall-based assessments, and teachers often have limited professional autonomy (Eurydice, 2025). The system prioritizes breadth over depth in subject matter, offering students a less specialized academic track compared to their European peers. In contrast, EU educational systems often feature early subject specialization, rely on essay-based

assessments and oral examinations that encourage deeper analysis, and adopt centralized curricula with consistent national standards (Au, 2011). Teachers in these systems typically hold advanced credentials and benefit from greater professional autonomy. In Germany, students specialize at an early age through distinct educational tracks, a structure that contributes to the country's consistently high performance on international assessments. France emphasizes a highly centralized national curriculum and essay-based examinations designed to promote analytical skills.

These pedagogical differences suggest that European education systems may have an edge over their American counterparts in cultivating better epistemic practices for some students (those in a science or professional tracks) but might not give repeated exposure to these types of lessons to students who focus on trade skills.

Media literacy initiatives in the post-truth era must go beyond identifying bias and evaluating sources (Tebaldi and NygreenView, 2021). Effective interventions in the U.S. and EU will take different forms, shaped by distinct cultural and institutional values. At the core of this divergence lies the tension between the values of liberty and dignity. In the U.S., a strong emphasis on individual sovereignty and the sanctity of one's personal informational "home" fosters greater distrust of government intrusion into media consumption. European societies, by contrast, place a stronger emphasis on public dignity and reputational integrity, often leading to greater suspicion of corporate influence over information systems.

These dynamics suggest that effective media literacy programs should be tailored to each region's sensibilities. In the U.S., where skepticism toward government is more pronounced, media literacy training might channel a more adversarial style—teaching individuals to challenge even their own side's narratives. In the EU, where citizens are generally more accepting of speech regulation, it may be more fruitful to focus on solidifying the mechanisms within the government or other trusted institutions that are meant to keep the systems honest and non-distorted.

In both contexts, however, the ultimate goal remains the same: to equip media consumers with the tools to see which forms of bullshit are likely to pull them off the track of justified belief, and in turn, to demand less of it.

6 Conclusion

Combating the spread of bullshit requires confronting an uncomfortable truth: its persistence is driven less by supply than by audience demand. Legal and policy efforts that focus primarily on curbing supply misdiagnose the problem. Bullshit thrives not because it deceives an unwilling public, but because it gratifies emotional, social, and identity-driven needs. Curbing bullshit, then, demands more than regulating its production. It requires a sustained effort to shift the preferences and incentives that fuel its consumption.

In the United States and the European Union, legal and policy responses aimed at tackling the problem diverge sharply, despite a shared emphasis on supply-side interventions. The U.S., constrained by strong constitutional speech protections, is limited in its ability to regulate content directly, and relies on limited regulatory tools and leans heavily on market forces and platform self-governance. The EU, by contrast, has embraced assertive top-down regulation aimed at reducing the visibility of harmful content. Yet neither approach can succeed without addressing the root of the problem: our shared human vulnerability to emotionally gratifying and self-affirming falsehoods.

Long-term progress depends not on banning bullshit, but on diminishing its appeal. In the U.S. this means relying on cultural, educational, and market-based strategies to cultivate a deeper public commitment to truth-seeking. In the EU, it requires complementing regulatory efforts with initiatives that promote critical thinking and media literacy. On both sides of the Atlantic, progress depends on sustained investment in demand side strategies capable of reshaping public informational habits and fostering a stronger preference for epistemic integrity.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Adams G, Bambauer J (2026) Curation as competition, curation as vice. Florida Law Review Forthcoming

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. Journal of Economic Perspectives 31(2):211–236

Arbel YA, Gilbert MD (2024) Truth bounties: A market solution to fake news. North Carolina Law Review 102:509

Armijo E (2021) Reasonableness as censorship. Florida Law Review 73:1199–1209

Au W (2011) Teaching under the new taylorism: High-stakes testing and the standardization of the 21st century curriculum. Journal of Curriculum Studies 43:25–45

Bloch-Wehba H (2026) Information law pluralism. Indiana Law Journal URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5163805, forthcoming, SSRN ID 5163805

Budak C, et al (2024) Misunderstanding the harms of online misinformation. Nature 630:45–46

Chuai Y, et al (2024) Did the roll-out of community notes reduce engagement with misinformation on x/twitter? Proceedings of the ACM on Human-Computer Interaction 8:1

EU (2000) Directive 2000/31/ec of the european parliament and of the council on certain legal aspects of information society services (e-commerce directive)

EU (2016) Regulation (eu) 2016/679 of the european parliament and of the council on the protection of natural persons (general data protection regulation)

EU (2022a) Regulation (eu) 2022/1925 of the european parliament and of the council on contestable and fair markets in the digital sector (digital markets act)

EU (2022b) Regulation (eu) 2022/2065 of the european parliament and of the council on a single market for digital services (digital services act)

Eurydice (2025) The structure of european education systems. URL https://eurydice.eacea.ec.europa.eu/data-and-visuals/european-education-structures

Florida (2021) Fla. stat. § 501.2041(2)(a) & (d)

Frankfurt H (2005) On Bullshit. Princeton University Press

Fukuyama F, Katz RR (2021) Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy. Hoover Institution

Gillespie T (2018) Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. London: Yale University Press

Grimmelmann J (2023) Listeners' choices online. URL https://james.grimmelmann.net/files/articles/listeners-choices-online.pdf, working draft

Guess AM, et al (2023) How do social media feed algorithms affect attitudes and behavior in an election campaign? Science 381(Special Issue). URL https://www.science.org/stoken/author-tokens/ST-1373/full

Hosseinmardi H, et al (2021) Examining the consumption of radical content on youtube. Proceedings of the National Academy of Sciences USA 118

Jones-Jang SM, et al (2019) Does media literacy help identification of fake news? information literacy helps, but other literacies don't. American Behavioral Scientist 2019:1–18

Kahneman D (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux

Langvardt K (2019) Regulating habit-forming technology. Fordham Law Review 88:129, 149, 154

Lin HS (2024) Towards implementation of warrant-based content self-moderation. Electronic Markets 34:1

New America (2023) By protecting data, federal privacy law can reduce algorithmic targeting and the spread of disinformation. URL https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/by-protecting-data-federal-privacy-law-can-reduce-algorithmic-targeting-and-the-spread-of-disinformation/

Nguyen CT (2020) Echo chambers and epistemic bubbles. Episteme 17:141–161

Norton H (2021) Manipulation and the first amendment. William & Mary Bill of Rights Journal 30:221

Roozenbeek J, van der Linden S (2019) Fake news game confers psychological resistance against online misinformation. Palgrave Communications 5:1

Rozenshtein AZ (2021) Moderating the fediverse: Content moderation on distributed social media. Journal of Free Speech Law 1:217–231

Salvi C, Barr N, Dunsmoor JE, et al (2023) Insight problem solving ability predicts reduced susceptibility to fake news, bullshit, and overclaiming. Thinking & Reasoning 29:760

Stray J, et al (2023) The algorithmic management of polarization and violence on social media. Tech. Rep. 23-05, Knight First Amendment Institute, URL https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media

Susser D, Roessler B, Nissenbaum H (2019) Online manipulation: Hidden influences in a digital world. Georgetown Law Technology Review 4:1, 26

Tebaldi C, NygreenView K (2021) Opening or impasse? critical media literacy pedagogy in a post-truth era. Cultural Studies Critical Methodologies 22:143–153

US C (2023a) Restrict act of 2023, s.686, §3. URL https://www.congress.gov/bill/118th-congress/senate-bill/686

US DC (2023b) Netchoice, llc v. bonta. URL https://cdn.ca9.uscourts.gov/datastore/opinions/2025/09/09/25-146.pdf

US SC (2024) Moody v. netchoice, 603 u.s. 707. URL https://www.supremecourt.gov/opinions/23pdf/22-277_d18f.pdf

Volokh E (2021) Treating social media platforms like common carriers? Journal of Free Speech Law 1:377

Wertenbroch K (1998) Consumption self-control by rationing purchase quantities of virtue and vice. Marketing Science 17:297

From Disinformation Dissection to Ideological Conflation

Leveraging Semantic Embeddings and Stochastic Processes for Detection of Behavioral Drifts Among Content-Creators

Clément Bénesse[1]

[1] Opsci.ai.

Corresponding author. E-mail: clement@opsci.ai

Abstract

"Magic-middle" social media influencers serve as critical intermediaries in contemporary information warfare, operating below systematic monitoring thresholds while maintaining trusted relationships within specialized niches. We present a mathematical framework for detecting coordinated influence operations through behavioural pattern analysis rather than content classification. Our approach models influencer digital footprints using semantic embeddings processed through hierarchical stochastic processes: Hidden Markov Models for inter-topic transitions and Ornstein-Uhlenbeck processes for intra-topic semantic drift. This content-agnostic methodology enables politically neutral detection. The framework provides anomaly detection through "cost of postage" metrics and structural comparison via Wasserstein distances between Gaussian processes.

Keywords

Disinformation Detection, Influencers, Embeddings, Behaviour Modeling, Stochastic Processes.

1 Introduction

In a polarized discourse, social networks become informational battlegrounds on which multiple actors push given narratives, either to discredit opponents or to muddy the waters and distract attention – part of what is now designated as hybrid wars (Bond, 2007). This mis- and disinformation is usually tailored under three main components: the factuality of the message (how true a given claim is, and how difficult it can be for readers to assess its truthfulness), its linguistic markers (rhetorical patterns, manipulative techniques in order to draw emotions, etc) and the communities relaying it (bots, astroturfing, complicit actors and organic growth). In such a context, influencers serve as key intermediaries, shaping how information spreads and reinforcing the dynamics of a two-step flow of communication (Viginum, 2025).

By holding a great deal of informational power in their niche, content-creators and influencers act as a trusted source of information for many individuals (Betts, Bliuc, and Courtney, 2025). As such, they can be valuable targets for propagandists, usually using the latter two elements of the described disinformation tryptic – e.g. recent reports by Viginum (Viginum, 2025) for empirical observations in France. To make the analysis more complicated, those actors do not exhibit drastic change of behaviour that would endanger their position of power in their niche but have the possibility to slowly drift towards different ideologies and affinities. This long-term, weak nudging is more insidious and harder to detect than drastic change of behaviour, making detection less clear.

We propose to leverage part of the NLP pipeline in what is a perfect example of how GenAI can be used as a defence against disinformation, while the opposite – AI enabling malicious actors – is usually shown as a common pitfall of this technology (Bontcheva, Papadopoulous, Tsalakanidou, Gallotti, Dutkiewicz, Krack, Teyssou, Severio Nucci, Spangenberg, Srba et al., 2024; Shoaib, Wang, Ahvanooey, and Zhao, 2023). From LLM usage for textual content to diffusion models for image, examples of synthetic content being used for manipulation of the public are common. However, the remedy may be in the poison as the same techniques can allow for wide improvements of fact-checkers and analysts workload, e.g. through better detection of malicious content or generation of prebunks.

As part of a wider project – the ENO-PROMPT project– where we dissect disinformation based on multiple prisms that gather specific insights and call AI agents for large scale analysis, we propose to dive into a mathematical framework in which influencers' digital footprint is modelled in semantic representations – embeddings (Muennighoff, Tazi, Magne, and Reimers, 2023) – by using stochastic processes (Cinlar, 2013; Helbing, 2010). Using this abstraction allows us to use classical out-of-distribution tools to identify a shift of behaviour of a given influencer, and if such exists to timestamp it.

To go even further, this methodology can be used to complement other tools for detection of coordination between influencers (Giglietto, Righetti, Rossi, and Marino, 2020; Wasserman, 1980), even if they have distinct communities and niches, or use different languages and platforms when using recent improvements in semantic embeddings. It can also be included as an additional agent, usable by automated pipelines commonly used by fact-checkers as these additional insights provide knowledge at a macro level on a critical category of actors to help analysts better tackle disinformation.

2 The "Magic Middle", or why study influencers behaviour with constraints

The detection of coordinated influence operations requires sophisticated analytical frameworks that can operate within the constraints faced by democratic institutions. A specific demographic of interest is the "magic middle" (Katz, 1940), a medium-size relay that holds a position of power on their informational niche and can rely content with implicit trust. Traditional approaches to analysis of such content often rely on explicit ideological classification or narrative categorization, which presents challenges for public entities bound by neutrality requirements and legal limitations on ideological profiling.

2.1 Background

The study of influence operations in digital spaces has evolved significantly with the proliferation of social media platforms. Traditional propaganda models, rooted in mass media theory, assumed centralized control and uniform message dissemination. However, contemporary information warfare operates through decentralized networks where influence is exerted through seemingly organic content creators who command devoted audiences within specific niches.

Research in computational social science has demonstrated that influence campaigns often target mid-tier content creators rather than mainstream media outlets or mega-influencers, see (Betts et al., 2025). This strategic choice stems from several factors: mid-tier influencers maintain higher engagement rates, possess more authentic relationships with their audiences, and operate with less scrutiny from platform moderation systems. Furthermore, their content appears more organic and less obviously sponsored, making manipulation attempts less detectable to both audiences and automated detection systems.

The challenge for researchers and analysts lies in distinguishing between genuine ideological evolution and coordinated influence operations. Natural drift in content creators' perspectives occurs regularly due to personal experiences, audience feedback, or changing social contexts. However, artificially induced shifts tend to exhibit specific mathematical signatures in timing, semantic consistency, and coordination patterns that can be detected through appropriate modelling frameworks.

2.2 The Magic Middle

Our framework operates on a "double magic middle" concept that targets two distinct but interconnected dimensions of influence operations. The first middle refers to mid-tier influencers – those with follower counts typically ranging from 10,000 to 500,000 across platforms. These actors occupy a sweet spot where they maintain significant reach while preserving authenticity and high engagement rates that mega-influencers often lose.

The second middle represents the semantic space between explicit propaganda and completely organic content. This is where the most sophisticated influence operations occur – through subtle shifts in framing, gradual introduction of new talking points, and strategic amplification of certain narratives while de-emphasizing others. Content in this middle ground maintains plausible deniability while systematically nudging audiences toward specific ideological positions.

This double targeting creates a multiplication effect: moderately-sized audiences receiving moderately-shifted content that appears authentic. The cumulative impact across multiple such actors can significantly influence public discourse without triggering obvious detection mechanisms. Our mathematical framework specifically addresses this challenge by identifying patterns that emerge only when analysing the collective behaviour of multiple actors over extended time periods.

2.3 Why not rely on narratology?

A straight-forward means to analyse the behaviour of given actors is through careful consideration of the content of their posts and the topic they choose to push. This field – which can be found under the name "narratology" and make use for instance of actantial theory (Elfes, 2025; Greimas, 2014) – investigates specifically the relationship and the ideal points of the narratives used in public discourse. While this approach exhibits strong results – and has been for a long time as it was first implemented for legacy media and political discourse – we choose deliberately not to make use of it.

One of the main reasons is that we want to explore a way, while using similar tools as explained in the next section, to assess deviations from an expected behaviour without relying on ideological cues such as polarization of the discourse. This is especially important for integration of those methodologies by public entities, as they usually have top-down or legal constraints that forbid them to do so.

Given such constraints, we therefore still rely on semantic spaces but we use them purely as black-boxes. Once a potential time interval of ideological conflation between actors is detected, our methodology stops and returns to the end user, usually a human analyst with expert knowledge on the constraints and needs, a window of interest to explore.

Note that in such a framework, one should always keep in mind the nature of the phenomenon at hand. In this case, it is the difference between legitimate mobilization and coordinated manipulations. The capacity of attribution remains an open question, especially considering that content pushed on social networks are always a fraction of the interaction between actors and therefore, cannot fully capture the full relationship between them. For this reason, our methodology can only flag behaviour of interest for later analysis and control by human experts. We emphasize that this tool cannot be used without due process by the down-stream analysts, who should consider and analyse outputs with their additional knowledge. Nonetheless, there are some insights that can help analysts:

1. Coordinated manipulation tend to happen in a short time-frame. Therefore, the time interval between pushed content is a precious indicator.

2. "Magic middle" influencers using their platforms as a relay tend to do it after commercial operations. Analysts with access to relevant informations on that topic can find it interesting for attribution. In that regard, a venue of future exploration is the literature regarding scam-detection on social networks.
3. Large jumps in the semantic space between usual topics of influencers and new discussions can be a flag that can be lowered by analysts considering additional communications by influencers, e.g. some real-life event that prompt an influencer to push specific content.

3 Mathematical modelling

In order to model an actor behaviour on social media, we embed the history of their posts using an embedding so that we can have a mathematically usable representation (Chen, Xiao, Zhang, Luo, Lian, and Liu, 2024; Muennighoff et al., 2023; Wang, Yang, Huang, Yang, Majumder, and Wei, 2024). Once this is done, we do not need the content any more. We can then aggregate the content by clusters that will be useful for our use of Markov Models. While these clusters may be representative of real-topics, we are not concerned with this point of view since their only use is to fit a model to assess how they may go from one topic to another. Once the inter-cluster dynamic is modelled, we focus on the intra-cluster dynamic by using stochastic processes. In particular, the Ornstein-Uhlenbeck process (Cinlar, 2013) is of special interest for us for its attraction / repulsion structure. By merging these two concepts, we obtain the final model that we fit according to the data obtained.

3.1 Embedding of influencers content

The first part of the pipeline is to embed the data in a semantic space. To do so, we use an embedding model, a commonly-used tool in the NLP literature. While using similar architectures to Large Language Models (LLMs) (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, 2023), they differ in the goal they aim to achieve. While LLMs are made to generate a new token given a series of context tokens – a task akin to time-series generation – embedding models are used to "put tokens that have similar semantic meaning close one to another". This means that they are well-adapted for tasks involving distance, leveraging the inner geometry of the model.

Additionally, newer generations of embeddings exhibit strong capabilities on various properties that were complicated to handle for previous generations. Beyond computational improvements, one can note that embedding models are now capable of handling multi-lingual settings – whereas content used to be concentrated based on used language – and recently provide some multimodal settings in which one can embed composite content – e.g. textual and video. When applied to a topic such as information on social media, where content is not unimodal and may be in a trans-national context, those improvements prove to be necessary to industrial adoption.

It is necessary to mention that one of the main tasks for which embeddings prove to be crucial is topic modelling, namely based on BERTopic methodology (Grootendorst, 2022). This task, aligned to our goal, aims to clusterize topics based on the ideological focal point they relay. Nonetheless, we re-iterate on the point that we start with the same pipeline but we then divert as we do not need to go back to the original content to infer some insights. Additionally, and before applying our stochastic modelling framework, we must mention that embedding models output in a high-dimensional space. Contemporary embedding models typically output vectors in spaces ranging from 384 to 1024 dimensions, which presents computational challenges for subsequent clustering and stochastic modelling. BERTopic addresses this through a dimensionality reduction step, typically employing UMAP (Uniform Manifold Approximation and Projection) (McInnes, Healy, and Melville, 2018) to project embeddings into lower-dimensional spaces of 5-6 dimensions, which we do too. Once one obtains the various clusters of interest given the history of content pushed by some actors, we focus on the two next steps: inter- and intra-topic modelling.

3.2 Inter-topic modelling

Given a dataset of content, sorted by timestamp, $\mathbf{X} = (x_1, \cdots, x_n)$ that we embed using a model f, we can use BERTopic – or any other methodology – to obtain a set of clusters $\mathbf{C} = (c_1, \cdots, c_j)$. The first step is to model how one goes from one cluster to another. The rationale behind this is that social media users, especially in the aforementioned magic middle, have some subjects that prove to gain traction with their followers. Therefore, they cycle through those themes. Knowing how they swap from one theme to

another allows to have a reference for later modelling. To do so, we use a Hidden Markov Model (Eddy, 1996), which makes the assumption that one only need to know the last topic in order to predict what will be the next topic of interest. Note that this allows us to obtain a stationary measure that simulates how users allocate their time across various subjects in the long run.

The estimation of the transition matrix, at the core of the Hidden Markov Model, can be done in a straightforward fashion by empirically computing the transition probability of jumping from cluster i to cluster j.

3.3 Intra-topic modelling

Given a cluster and for intra-topic dynamics, we rely on stochastic processes. The rationale behind this choice is that, having a cluster, we have the possibility to have a centroid which would the core idea – or focal point – of the theme at hand to which we want to have some form of attraction, and a perimeter – in other words, some form of repulsion from the centroid that we can still quantify. In order to do this, we use the Ornstein-Uhlenbeck (Maller, Müller, and Szimayer, 2009) model defined by the following equation:

$$dx_{j,t} = -\theta_j(x_{j,t} - \mu_j)dt + \sigma_j dW_{j,t}, \tag{1}$$

where $x_{j,t}$ is the content in cluster j at time t, θ_j, μ_j and σ_j are parameters to be fitted (μ_j being the mean-reversion or "centroid of the cluster" in the semantic space) and $W_{j,t}$ is a Weiner process.

Note that this process is especially interesting for us since it is in fact a stationary Gauss–Markov process, which means that it is a Gaussian process (MacKay et al., 1998), a Markov process, and is temporally homogeneous. This is useful for several reasons. First, its interaction with the Hidden Markov Model means that we only need to keep the last state of each cluster in order to simulate from this process. Second, we have closed formulae, see (Maller et al., 2009), for the parameters and to update them, leveraging the Gaussian process literature. Finally, being Gaussian means that some distributional distances are easily computed in order to compare the stationary processes for two given influencers, e.g. the Wasserstein distance as we will see later.

4 Outputs and Methodology

This combined framework yields several analytical capabilities that enable detection of coordinated behaviour patterns while maintaining the political neutrality essential for institutional deployment. The structural modelling provide objective and intrinsic measures of behavioural deviation and coordination that do not require explicit ideological labelling or content interpretation. We present three key applications of this framework: anomaly detection through cost-of-postage metrics, comparative analysis via structural differences, and coordination detection through ideological conflation analysis.

4.1 Cost of postage

The "cost of postage" metric quantifies how surprising a given piece of content is relative to an influencer's established behavioural model. By computing the negative log-likelihood of new content under the fitted mixed Gaussian process, we obtain a score that reflects semantic departure from expected patterns.

Formally, for a new message x_{new} in cluster j at time t, the cost of postage can be defined as the log-likelihood of the density of the fitted process $\hat{\mathbb{P}}(x_{new}) = \sum \lambda_j \mathbb{P}_j(x_{new}, \hat{\theta}_j, \mu_j, \sigma_j,)$, where λ_j is the weight of the cluster j for the stationary measure of the HMM model and the parameters $\theta_j, \mu_j, \sigma_j$ are for the Ornstein-Uhlenbeck process of the cluster j.

High cost of postage scores indicate content that deviates significantly from the influencer's historical patterns, potentially signalling external influence or coordinated messaging.

Note that another element of interest using $\hat{P}$ is that we can use it as a generator, sampling potential datapoints. A concrete use case is the development of scenarii in order to mimic personae, for instance for PR training.

4.2 Structural differences

Since we obtain Gaussian Processes, we can leverage the fact that the 2-Wasserstein distance has closed formulae in that case (Villani, 2021). Note that these formulae only require comparison of the mean and the variance of the processes, making computation efficient and allowing for online implementation – prone to addition of newly found data.

For two influencers A and B with respective mixed Gaussian processes characterized by parameters $(\theta_j^A, \mu_j^A, \sigma_j^A)$ and $(\theta_j^B, \mu_j^B, \sigma_j^B)$ across clusters, we compute structural similarity at multiple levels. First, we compare their HMM transition matrices P^A and P^B using the Frobenius norm $||P^A - P^B||_F$, which captures differences in how they navigate between topics. Second, for each shared cluster j, we compute the Wasserstein distance between the stationary distributions of their respective Ornstein-Uhlenbeck processes:

$$W_2^j(A, B) = ||\mu_j^A - \mu_j^B||^2 + \text{tr}(\Sigma_j^A + \Sigma_j^B - 2((\Sigma_j^A)^{1/2}\Sigma_j^B(\Sigma_j^A)^{1/2})^{1/2}) \tag{2}$$

where Σ_j^k represents the covariance matrix of the stationary distribution for influencer k in cluster j. The overall structural distance combines these measures :

$$D_{struct}(A, B) = \alpha||P^A - P^B||_F + \beta \sum_j W_2^j(A, B) \tag{3}$$

where α, β are scaling parameters. This metric enables identification of influencers with similar behavioural patterns even when their surface content appears different, which is crucial for detecting coordinated campaigns that maintain plausible deniability through topic diversity.

4.3 Ideological conflation

By integrating the time component, analysts and end-users can get access to specific time intervals in which we detect potential ideological conflation. By ideological conflation, we aim to detect the phenomenon in which a malicious actor that want to propagate disinformation or false narratives regarding a given ecosystem, and does so by pushing this content through the relay of multiple influencers. By economical pragmatism and reach optimization, the typology of influencers that offer the most impact and reception to their audience, while remaining financially accessible, are the influencers of medium size, which correspond to the "magic middle" also defined in the digital marketing literature. Therefore, we target abnormal behaviours from multiple actors that are concentrated around similar timestamps. There are mainly two ways to detect such a phenomenon.

1. The first approach is to use the complete modelling made earlier to leverage the cost of postage metric. By doing so, we flag as suspicious content that exhibit proximity in the embedding space and that have high cost of postage, at least for one actor. Note that if the cost of postage is high only for one actor (e.g. A), that relates to opinion shifting from A towards the topic of B. If both actors exhibit high cost of postage, then the topic at hand is unfamiliar for A and B, which may hint at exogenous influence.
2. The second approach is more compute-intensive. It revolves around adopting a rolling time window and to fit processes for each of time bucket. Thanks to this scheme, and by using the D_{struct} metric mentioned above, one can plot across time the similarity between these actors, leading to identification of conflation by looking at minima from the D_{struct} metric.

For both of those methods, there is a need to determine a threshold that trigger alerting. We recommand using threshold based on quantiles of the distribution of dedicated metrics. Indeed, it is complex to determine a specific value, as range depends in some part on the volume of data available, on the practice of each influencer, etc.

5 Discussion

In this work, we considered a mathematical representation of influencers behaviour by making use of stochastic processes and embedding inner semantic representations. This mathematical framework represents an interesting advancement in automated detection of influence operations targeting social media creators. By abstracting away from content specifics while preserving semantic relationships, the approach enables detection capabilities that are both politically neutral and linguistically agnostic. This addresses limitations of existing approaches that often require manual specification of ideological markers or language-specific processing pipelines.

5.1 Limitations

However, several limitations remain. We recognize that our approach is not entirely exempt of potential biases. We rely on the choice of an embedding model, along with a dimension reduction and clustering

methodology that contain thresholds that should be set by analysts. As any ML model, embeddings encode a representation of society perceived through the prism of training data and therefore can contain latent biases that can have an impact on the tool. This is mainly explored by the explainability and interpretability literature (Jourdan, Picard, Fel, Risser, Loubes, and Asher, 2023; Weber, 2025), with several open questions in that regard. We once again emphasize that such choices, while allowing analysts to map dynamics across multilingual settings and therefore catching phenomenon that would require different skills otherwise, should not be used in a techno-solutionist approach that forego institutional oversight.

Additionally, the use of embedding models means that our methodology can only be as good as this brick. If a model is not expressive enough, then our methodology will be faulty due to that aspect. This means that the tool should be considered through the lens of multilingual capabilities of the embedding models, for which multiple benchmarks exist, see (Enevoldsen, Chung, Kerboua, Kardos, Mathur, Stap, Gala, Siblini, Krzemiński, Winata et al., 2025). Low-resources languages should be carefully examinated as they are usually domains in which embedding models can be less efficient and therefore could lead to increased failure rates. Nonetheless, this issue improves as newer-generation models prove to be efficient and to provide meaningful representations. Finally, the framework requires sufficient historical data to establish reliable baseline models, limiting its effectiveness for newly emerged influencers or rapidly evolving topics.

5.2 Perspectives

Our future work will explore the empirical aspect of this methodology, focusing on real-life datasets and simulated examples. It will included guardrails and recommendations for hyper-parameter choices, and discuss governance frameworks for interested public institutions that want to integrate this methodology in their technical toolbox.

Additionally, we are interested in exploring adaptive modelling approaches that can accommodate changing content landscapes, integration with multimodal embedding models to capture visual and audio content as those format become prevalent on social media, and development of real-time alerting systems suitable for operational deployment by fact-checking organizations.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Betts JM, Bliuc AM, Courtney DS (2025) The role of charismatic influencers in online polarisation: An agent-based modelling approach. Available at SSRN 5170258

Bond MS (2007) Hybrid war: a new paradigm for stability operations in failing states

Bontcheva K, Papadopoulous S, Tsalakanidou F, et al (2024) Generative ai and disinformation: recent advances, challenges, and opportunities

Chen J, Xiao S, Zhang P, et al (2024) Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. URL https://arxiv.org/abs/2402.03216, arXiv:2402.03216

Cinlar E (2013) Introduction to stochastic processes. Courier Corporation

Eddy SR (1996) Hidden markov models. Current opinion in structural biology 6(3):361–365

Elfes J (2025) Mapping news narratives using llms and narrative-structured text embeddings. In: Proceedings of the 17th ACM Web Science Conference 2025, pp 326–337

Enevoldsen K, Chung I, Kerboua I, et al (2025) Mmteb: Massive multilingual text embedding benchmark. arXiv preprint arXiv:250213595

Giglietto F, Righetti N, Rossi L, et al (2020) It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 italian elections. Information, Communication & Society 23(6):867–891

Greimas AJ (2014) Reflections on actantial models. In: Narratology. Routledge, p 76–89

Grootendorst M (2022) Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:220305794

Helbing D (2010) Quantitative sociodynamics: stochastic methods and models of social interaction processes. Springer Science & Business Media

Jourdan F, Picard A, Fel T, et al (2023) Cockatiel: Continuous concept ranked attribution with interpretable elements for explaining neural net classifiers on nlp tasks. arXiv preprint arXiv:230506754

Katz E (1940) The two-step flow of communication: An up-to-date report. Public Opinion Quarterly 21

MacKay DJ, et al (1998) Introduction to gaussian processes. NATO ASI series F computer and systems sciences 168:133–166

Maller RA, Müller G, Szimayer A (2009) Ornstein–uhlenbeck processes and extensions. Handbook of financial time series pp 421–437

McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426

Muennighoff N, Tazi N, Magne L, et al (2023) Mteb: Massive text embedding benchmark. URL https://arxiv.org/abs/2210.07316, arXiv:2210.07316

Shoaib MR, Wang Z, Ahvanooey MT, et al (2023) Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large ai models. In: 2023 international conference on computer and applications (ICCA), IEEE, pp 1–7

Vaswani A, Shazeer N, Parmar N, et al (2023) Attention is all you need. URL https://arxiv.org/abs/1706.03762, arXiv:1706.03762

Viginum (2025) War in Ukraine: three years of Russian information operations

Villani C (2021) Topics in optimal transportation, vol 58. American Mathematical Soc.

Wang L, Yang N, Huang X, et al (2024) Multilingual e5 text embeddings: A technical report. URL https://arxiv.org/abs/2402.05672, arXiv:2402.05672

Wasserman S (1980) Analyzing social networks as stochastic processes. Journal of the American statistical association 75(370):280–294

Weber T (2025) Explainability for embedding ai: aspirations and actuality. arXiv preprint arXiv:250414631

Who Controls the Narrative? The Dual Role of Contributors and Amplifiers in Online Activism

Soo Young Bae[1], David DeFranza[2]

[1] Communication, University of Massachusetts Amherst.
[2] Marketing, University College Dublin.

Corresponding authors. E-mails: sooyoungbae@umass.edu; david.defranza@ucd.ie

Abstract

Hashtag activism has fundamentally transformed how social movements are organized and expressed through online social networks. Yet, as social movements gain visibility in digital platforms, they also become susceptible to politicization and discursive contestation. Our study focuses on this potential of social movements to be politicized and investigates how different types of participants contribute to the trajectory and meaning of hashtag-based activism in social media. Drawing on a large dataset of over one million tweets in the United States collected during the peak months of the #MeToo movement, we examine how users of different types utilize the hashtag #MeToo over time. In addition, we conduct a semantic network analysis to understand communities that reflect divergent framing strategies. Our results show that users who contribute original content to the #MeToo discourse show different patterns of hashtag usage in comparison to users who primarily amplify existing content. Additional semantic network analyses of user communities also suggest that reframing of the movement within broader ideological battles occur. We discuss the implications of these discursive dynamics in shaping the narrative of social movements.

Keywords

Social Media Activism, Hashtag Propagation, Narrative Control, Digital Influence, Semantic Network Analysis.

1 Introduction

The rise of social media has fundamentally reshaped the architecture of collective action, giving rise to what is often termed "hashtag activism." Enabled by the affordances of online social networks, citizens can now participate in social movements by circulating hashtags that bring visibility to social issues and injustices (Tufekci, 2017). Unlike traditional protests that rely on formal leadership structures and hierarchical organizations, digital movements frequently emerge organically and gain momentum through the decentralized activity of individual users. In this context, hashtags serve both as organizational tools and as symbolic frames through which public discourse is constructed and amplified.

Hashtag activism has proven especially powerful in elevating issues that were previously marginalized or stigmatized. The #MeToo movement illustrates how social media can function as a tool for mobilization and a discursive arena. Initially created as a call to share personal experiences of sexual harassment and abuse, #MeToo quickly evolved into a global conversation about gender, power, and inequality. The openness that makes social media an accessible platform for discourse development and storytelling also renders it vulnerable to discursive contestation. As social movements gain visibility, they invite not only support but also backlash, reinterpretation, and polarization. Prior research has shown that as hashtags circulate more widely, they often become sites of political struggle. Counter-movements such as #AllLivesMatter emerged in direct response to #BlackLivesMatter, reframing the original message to suggest exclusion or bias (Jackson, Bailey, and Welles, 2020). Similarly, in a study of the #MeToo movement in South Korea found that polarized political rhetoric frequently accompanied the hashtag (Bae, Kim, Ha, and Cha, 2021). Some users were shown to leverage the visibility of the movement to promote partisan agendas, redirect the narrative, or even cast doubt on the credibility of survivors.

We argue that such contestation reflects the deeper dynamics of narrative control in the digital public sphere. While some participants use hashtags to advance the original cause, others may seek to reframe the conversation to align with their ideological preferences. Such differences may occur not only between supporters and non-supporters of the movement, but also among different types of participants within the movement itself. This study draws attention on the communicative roles individuals assume based on their discursive action. Specifically, we examine the #MeToo movement on Twitter (now X) to understand how "contributors" (users who generate original content) and "amplifiers" (users who primarily retweet or share existing content) engage with the hashtag and influence its narrative trajectory. By categorizing the tweets according to the users' communicative behavior, we aim to provide a more granular understanding of how narrative control is distributed and contested within digital activism.

Our approach builds on existing literature that recognizes the heterogeneity of participation in online movements (Freelon, McIlwain, and Clark, 2016). While prior studies have examined aggregate patterns of hashtag use or shifts in framing over time, few have investigated how different types of users contribute to these patterns. We argue that this distinction between contributors and amplifiers provides a valuable perspective for understanding the complex narrative structure of online social movements.

2 Theory Development

2.1 Social Movements in the Digital Age

Online social networks have transformed the meaning of social movements and collective action. Rather than relying on formal organizations, many movements today emerge and evolve within digital spaces, where visibility and momentum are often measured by the virality of a hashtag. Numerous cases around the world have shown the power of networked activism, which often relied on loosely connected publics that mobilize around a shared hashtag. Although online hashtag movements, such as #BlackLivesMatter and #MeToo were rooted in offline events and lived experiences, hashtags served as entry points for constituents to participate, enabling individuals to contribute their voices and allowing disparate experiences to be connected into a recognizable social narrative (Jackson et al., 2020).

What distinguishes hashtag activism on social media is not only the speed and scale at which it can unfold, but also its ability to blur the boundaries between the personal and the political, and the public and the private. This blurring has important implications for how frames are constructed in digital environments, as compared to traditional media. Framing theory suggests that the media not only defines what the problem is, but also attributes responsibility and suggest remedies to the problem (Entman, 1993; Iyengar, 1990). In the context of social media, however, framing is not solely controlled

by journalists or institutional actors, and it can thus become a participatory and contested process. Users actively contribute to shaping the frame, thereby redistributing power and dominant narratives. Such openness makes digital movements vulnerable to discursive challenges.

Indeed, research has shown that social media platforms often become sites of oppositional framing, where movements are reframed to undermine their legitimacy or redirect their message. For instance, counter-narratives like #AllLivesMatter emerged not as extensions of #BlackLivesMatter, but as deliberate efforts to dilute its political meaning (Jackson et al., 2020). Similarly, polarized rhetoric frequently accompanied the #MeToo movement in South Korea, as some users exploited the hashtag's visibility to advance partisan agendas or discredit survivors (Bae et al., 2021). This means that the same digital infrastructures that enable grassroots visibility and participation can also facilitate the strategic hijacking of movements by political actors or ideological agendas (Jackson et al., 2020; Liu and McLeod, 2019; Nikolayenko, 2019). As a result, understanding how a movement's narrative takes shape on social media requires attention not only to the content of the discourse but also to the roles and intentions of those participating in it. To unpack this process, we focus on the varied communicative roles that individuals adopt when engaging with social movements online.

2.2 Contributors, Amplifiers, and the Dynamics of Hashtag Activism

The participants in hashtag activism engage in narrative shaping process through distinct communicative actions, primarily enabled by the different affordances of social media platforms. While some users function as "contributors," generating original posts that include personal narratives or commentary, others act as "amplifiers," primarily circulating content through retweets, shares, or endorsements without introducing new perspectives. We find that these different actions reflect more than communicative preferences, they signal different motivations for engagement.

The uses and gratifications theory conceptualizes media engagement as a goal-driven activity oriented toward fulfilling distinct communicative needs—ranging from informational and emotional to social and recreational (Katz, Blumler, and Gurevitch, 1973). In the context of online activism, this framework provides a valuable lens for understanding variation in user engagement. A contributor to the #MeToo movement might be motivated by the need for self-expression or solidarity building, while an amplifier might be driven by a desire to align with a message and gain visibility by supporting a resonant narrative. Although the amplification process itself plays a critical role in expanding the movement's reach and visibility, it can also subtly redirect or reframe the original message by selectively promoting content that aligns with distinct ideological or political goals (Freelon et al., 2016). This interplay between content creation and circulation underscores the dual nature of digital activism as both a collaborative and contested process of narrative building and control.

Recognizing the distinct roles that participants play in shaping online discourse around hashtag activism, we focus our attention on how social movement narratives evolve in digital spaces. Building on prior research that highlights the susceptibility of online movements to discursive contestation and political appropriation (Bae et al., 2021; Jackson et al., 2020), we examine how contributors and amplifiers engage with the movement in distinct ways. To investigate these dynamics, we pose the following research questions:

- How do contributors and amplifiers **differ** in their use of the #MeToo hashtag **over time**?
- Do these groups form **distinct semantic communities**, reflecting divergent **framing strategies**?

3 Methods

3.1 Data Collection and Annotation

The study utilized a Twitter dataset procured from a third-party aggregator comprising 1,638,635 tweets from 39,112 users. The data was collected between September 24, 2017, and December 31, 2017, using the query ("metoo" OR "#metoo") and was restricted to tweets in English that originated in the United States.

A multi-stage annotation process was employed to generate labels for model training. To create user type classifications, a subsample of 1,000 tweets was randomly selected from the main dataset. Each research team member independently labeled 500 tweets as being an exemplar of either an "amplifier" or a "contributor" post. Criteria established before labeling defined contributor tweets as those including original thoughts or commentary, and amplifier tweets as those which either shared a link or retweeted

without the addition of original commentary. Labels were applied at the level of the tweet, with the positive label 1 used to indicate a contributor and the negative label 0 used to indicate an amplifier. Having established labeling criteria and applying labels to the subsample, the team reviewed all labels collectively. Any disagreements were resolved through discussion to produce a final verified set of 1,000 labeled tweets.

To categorize hashtag themes, all hashtags were extracted from the corpus and counted. High-frequency hashtags were defined as those with a frequency count three or more standard deviations above the mean, resulting in a selection of 119 unique hashtags for analysis. These hashtags were then collaboratively labeled by the research team as either MeToo-relevant, politics-relevant, or not relevant, with any discrepancies resolved through discussion.

3.2 Predictive Modeling Pipeline

A machine learning pipeline was applied to classify the full dataset of tweets. A minimal preprocessing strategy was first applied, which involved transforming all emojis into text descriptors and converting all URLs into a singular URL indicator token. Following this, we extracted dense feature representations from the tweet text using BERTweet, a transformer-based model pretrained on English-language Twitter data that has been shown to outperform other benchmarks on social media data (Nguyen, Vu, and Nguyen, 2020).

For the classification task, we employed an XGBoost model, a scalable tree-boosting system known for its ability to handle non-linearities and multicollinearity (Chen and Guestrin, 2016). This algorithm sequentially builds decision trees, with each new tree trained to correct the errors of its predecessor. The model inherently limits overfitting by penalizing unnecessary complexity and limiting tree size, and it has demonstrated superior performance compared to other methods, including deep neural networks, in various applications (Rafieian, 2023; Smith, Seiler, and Aggarwal, 2023).

The model's hyperparameters were tuned using a random search protocol with 30 iterations of 10-fold cross-validation performed on the labeled training set. The final hyperparameters were set as follows: a subsample rate of 0.7, which introduces randomness for better generalization by training each tree on 70% of the data; a regularization lambda of 1.0 to apply moderate L2 regularization and prevent overfitting; and the number of estimators set to 200, indicating the model would build 200 sequential trees. Furthermore, the maximum depth for each tree was limited to 3 to capture simpler patterns; the learning rate was set to a moderate value of 0.1 to control the contribution of each tree; gamma was set to 0, meaning splits were allowed if they reduced error even slightly; and the number of randomly sampled columns or features was set to 0.9, randomly selecting 90% of features for each tree to reduce overfitting. The fully trained model was then used to classify the entire unlabeled dataset.

3.3 Statistical and Network Analyses

To analyze temporal dynamics in the online movement, we aggregated tweet data into daily hashtag counts over the 98-day observation period, a technique used in prior literature to create a more stable unit of analysis (Bhatia and Walasek, 2016; Snefjella and Kuperman, 2015). Each count was log-transformed as $\log(count + 1)$ to improve interpretability. An ordinary least squares (OLS) regression model was fit to this data with the specification:

$$\log(count + 1) = \alpha + \beta\text{FOCUS} + \beta\text{CLASS} + \beta\text{TIME} + \beta\text{FOCUS} \times \text{CLASS} \times \text{TIME} + \epsilon$$

where FOCUS denotes the hashtag category (MeToo or politics), CLASS is the user type (amplifier or contributor), and TIME represents the day of the observation. β are parameters to be estimated and ϵ is the residual or error of the model.

Finally, to investigate the thematic structure of the discourse, we conducted a semantic network analysis. We constructed separate networks for amplifiers and contributors based on the co-occurrence of the 119 most frequent hashtags. To improve interpretability, the initial networks, containing 383 edges for amplifiers and 413 for contributors, were pruned to retain only connections in the upper quartile of co-occurrence frequency, resulting in final networks of 88 and 99 edges, respectively. We identified sub-communities within these networks using the Louvain community detection algorithm, a method designed to find subgraphs with high modularity (Blondel, Guillaume, Lambiotte, and Lefebvre, 2008; Li, Qin, Jiang, and Pedrycz, 2021; Traag, Waltman, and Van Eck, 2019). Finally, edges were curved

and bundled using the kernel density bundling algorithm to enhance interpretability (Hurter, Ersoy, and Telea, 2012).

4 Results

4.1 Classifaction Model Performance

The final model achieved strong classification performance, as indicated by an accuracy of 0.872, indicating that 87% of all predictions were correct as compared to the ground truth label; precision of 0.858, indicating that when the model predicted a positive label, it was correct 85.8% of the time; recall of 0.842, indicating that the model was able to capture 84% of all actual positive cases; a F1 score of 0.85, reflecting the prior evaluation result that the model was able to capture true positive cases without making false positive errors; and an ROC-AUC score of 0.935, indicating excellent performance in separating positive and negative classes (see Figure 1).

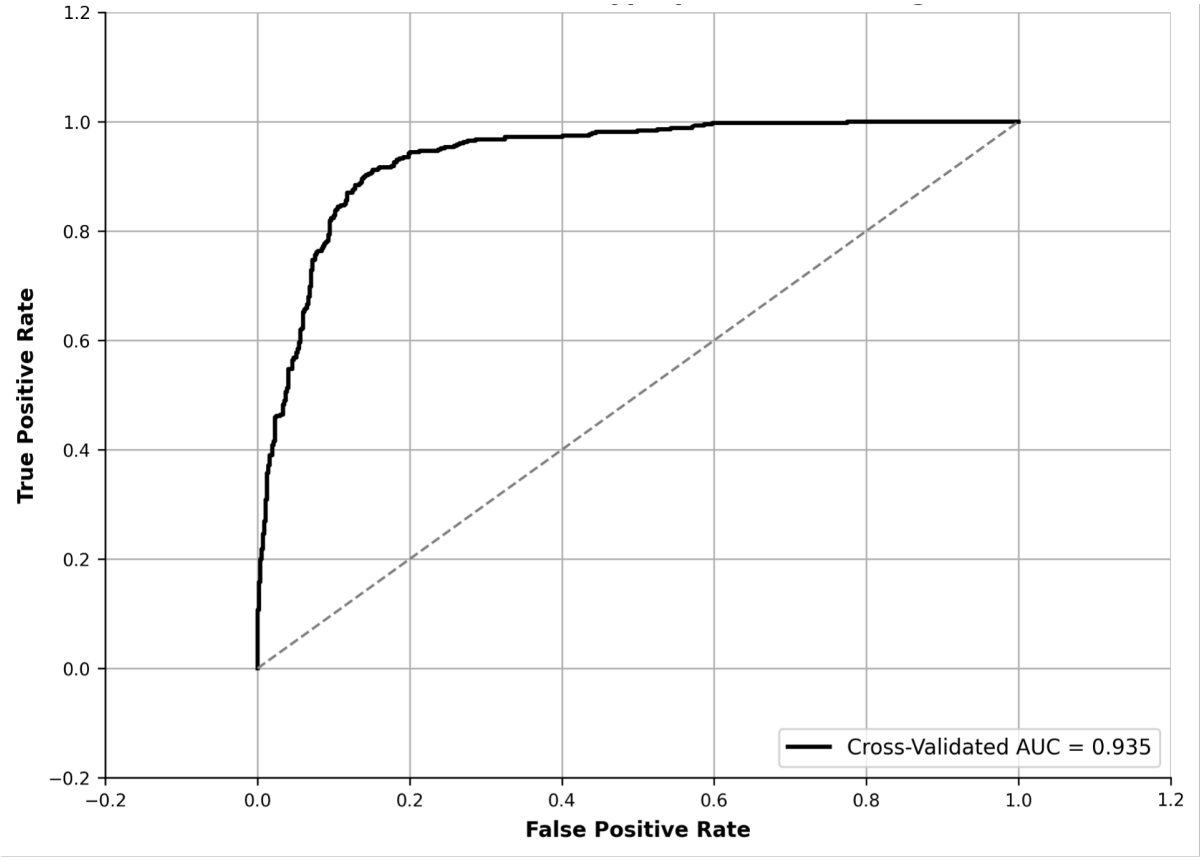


Fig. 1: ROC curve of the trained model after hyperparameter tuning.

4.2 Temporal Trends in Hashtag Use

The OLS regression model revealed significant interactions between user class, hashtag focus, and time. The overall model was significant, $F(7, 230480) = 43,960, p < 0.001$, and accounted for 57.2% of the variance in hashtag counts ($R^2 = 0.5718$). All model coefficients were statistically significant ($p < 0.001$), including the critical three-way interaction term (FOCUS × CLASS × TIME), as shown in Table 1.

The predicted marginal effects, visualized in Figure 2, illustrate these complex dynamics. Initially, amplifiers posted a higher volume of both MeToo-relevant and politics-relevant hashtags compared to contributors. Over the 98-day observation period, however, their behavior changed dramatically: amplifiers significantly decreased their use of MeToo hashtags while substantially increasing their use of general politics hashtags. Contributors also showed a decrease in MeToo hashtags and an increase in political hashtags, but these changes were far less pronounced. This finding indicates that, compared

Variable	β (SE)
Intercept	7.627^{***}
	(0.0061)
FOCUS (Politics)	-2.4809^{***}
	(0.0087)
CLASS (Contributor)	-2.0389^{***}
	(0.0145)
TIME	-0.0359^{***}
	(0.0001)
FOCUS (Politics) × CLASS (Contributor)	0.6873^{***}
	(0.0205)
FOCUS (Politics) × TIME	0.0675^{***}
	(0.0002))
CLASS (Contributor) × TIME	0.0257^{***}
	(0.0003)
FOCUS (Politics) × CLASS (Contributor) × TIME	-0.037^{***}
	(0.0004)
R^2	0.5718
Adjusted R^2	0.5718
$F(7, 230480)$	$43,960^{***}$

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 1: Results of the ordinary least squares regression of hashtag use over time.

to amplifiers, contributors maintained a more thematically consistent focus on the #MeToo movement over time.

4.3 Semantic Network Topology

The semantic network analysis suggested fundamental differences in the discourse structure of amplifiers and contributors. The Louvain algorithm identified seven unique communities in the amplifier network and six in the contributor network. A comparative analysis of the two largest communities from each group revealed divergent thematic priorities. In interpreting the network topologies, it is worth noting that the relative size of each node represents the frequency of occurrence for that hashtag in the group of tweets. By virtue of the kernel density bundling algorithm (Hurter et al., 2012), the relative thickness of the edge indicates a stronger cooccurrence of two connected hashtags, while the relative distance indicates the relatedness (mean path length) of the connected hashtags.

For amplifiers, general politics hashtags such as #impeachtrump, #trump, and #maga consistently occupied the most central positions in the largest communities (see Figure 3). Hashtags directly related to the #MeToo movement, such as #sexualassault and #harveyweinstein, were present but held more peripheral, dependent positions (see Figure 4). This structure suggests that the amplifier discourse assimilated the #MeToo movement into a broader, overarching political narrative.

In contrast, the contributor group displayed a more varied pattern. While the largest community was also centered on political hashtags (see Figure 5), the second largest community was almost exclusively saturated with MeToo-relevant hashtags (see Figure 6). In this latter community, terms such as #sexualharassment, #sexualassault, #womensrights, and #harveyweinstein formed a dense, central cluster, indicating that a significant portion of the contributor discourse was narrowly focused on the #MeToo movement itself, independent of a dominant political frame.

Taken together, these results reveal a clear divergence in the behavior and thematic focus of two distinct user groups within #MeToo discourse on Twitter. The analysis distinguished between "amplifiers," who initially engaged with the movement but quickly pivoted to a broader political narrative, and "contributors," who maintained a more sustained and focused conversation on #MeToo-specific topics. The temporal analysis demonstrates that while amplifiers' interest in #MeToo hashtags waned and was replaced by a sharp increase in political hashtagging, contributors' engagement remained more thematically consistent over time. This distinction was supported by the semantic network analysis,

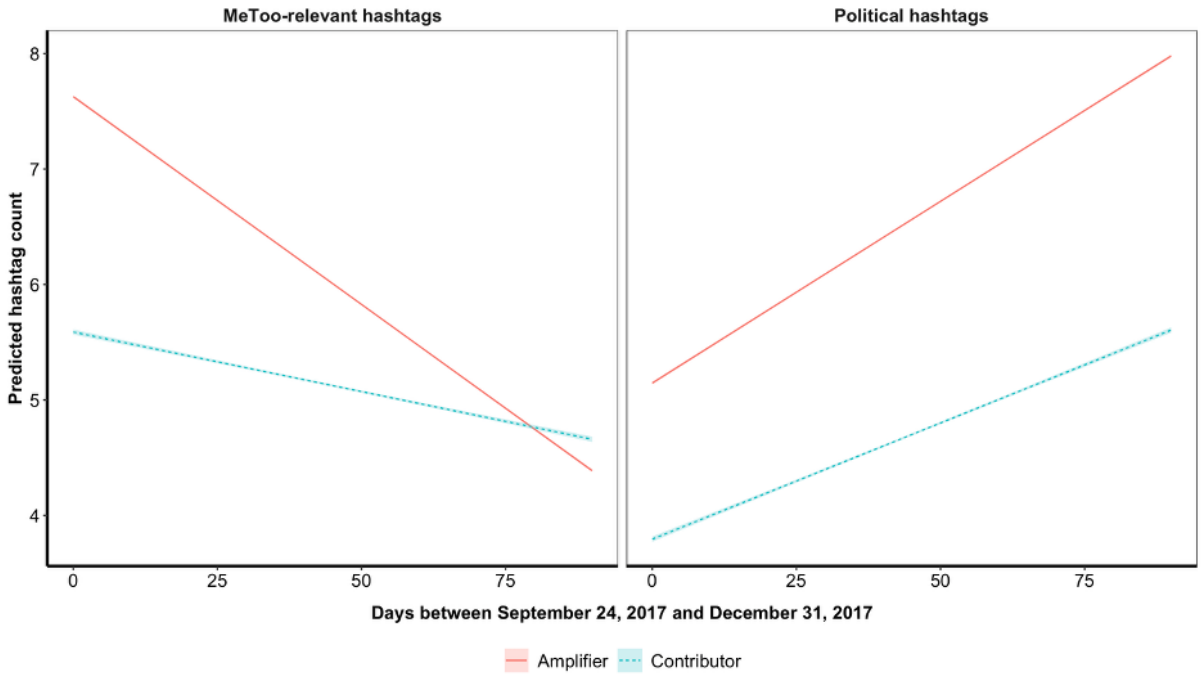


Fig. 2: Predicted marginal (absolute) effects of the change in hashtag use by amplifiers and contributors over time. Note: Shaded regions represent the 95% confidence bounds.

which showed that amplifier discourse was structurally dependent on a core of general political hashtags, whereas contributor discourse contained a significant, self-contained community focused almost exclusively on the vocabulary of the #MeToo movement itself. These findings suggest that large-scale online social movements are not monolithic; rather, they are composed of distinct user roles with divergent communicative patterns, motivations, and narrative trajectories that can be effectively identified through a combination of predictive modeling and network science.

5 Conclusion

This study examined how different types of participants, contributors and amplifiers, engaged with the #MeToo movement on Twitter during its early months of development. Our analysis revealed a clear divergence in framing practices between these two groups. Over the 98-day observation period, contributors were significantly more aligned with hashtags directly related to #MeToo (e.g. #whyistayed, #sexualharassment, #domesticviolence), while amplifiers were more likely to introduce and associate political hashtags with the movement (e.g. #resist, #impeachtrump, #notmypresident). This finding supports the view that the discourse on social movements is shaped by a multiplicity of actors who bring distinct communicative intentions to the conversation.

The fact that contributors maintained a consistent focus on movement-specific content suggests a commitment to the original goals of #MeToo. Such contributors shared personal testimony and raised awareness of relevant issues. These users functioned as narrative anchors, reinforcing the core framing of the movement through their experiences and opinions. This is consistent with findings from previous studies exploring online activism which highlight the role of personal narratives as a foundation of its discourse on social media (Clark, 2016).

In contrast, amplifiers appeared to integrate broader political discourses into the #MeToo movement significantly more than contributors. While amplifiers can also effectively expand the reach of a movement through creating a greater echo to the voices of contributors, this behavior also introduces the possibility of reframing or hijacking of the hashtag movement in a manner that reflects different motifs. By embedding #MeToo within political narratives, amplifiers may inadvertently or intentionally redirect the core focus of the movement. This pattern resonates with the literature on discursive hijacking, where visibility opens a movement to competing interpretations and strategic repurposing (Jackson et al., 2020). In this context, our study underscores that amplifiers may play a dual role by increasing

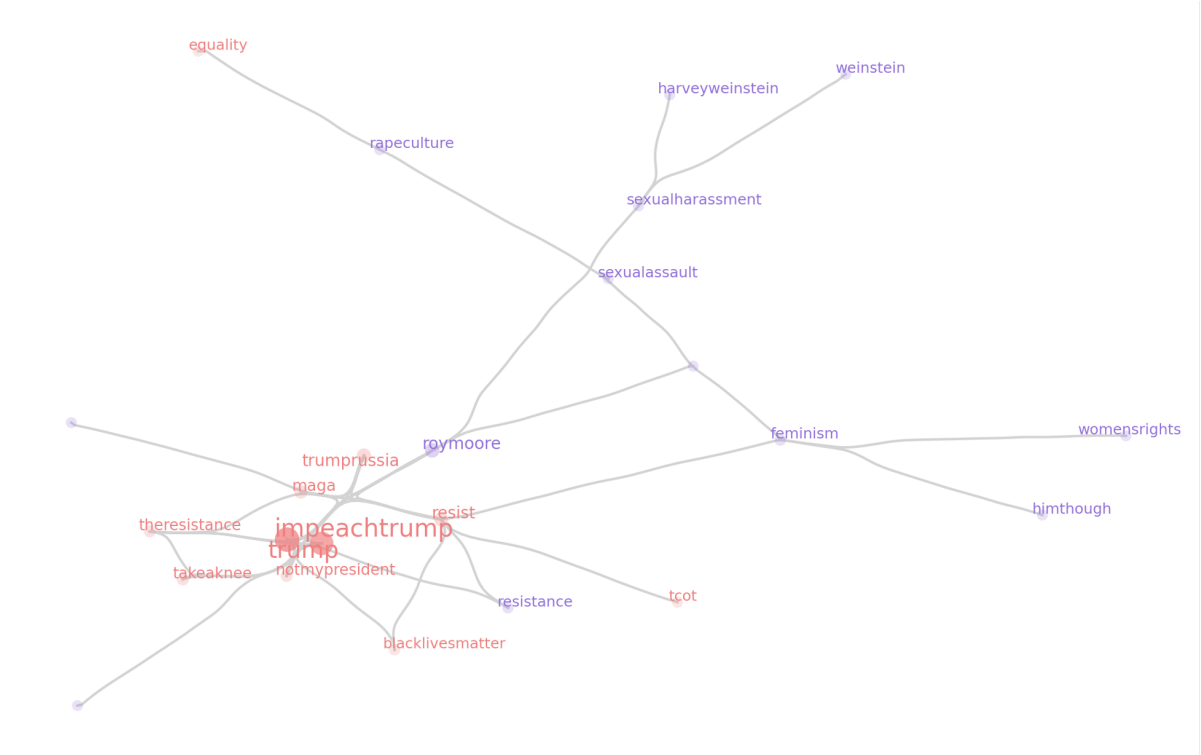


Fig. 3: The largest hashtag community among tweet amplifiers.

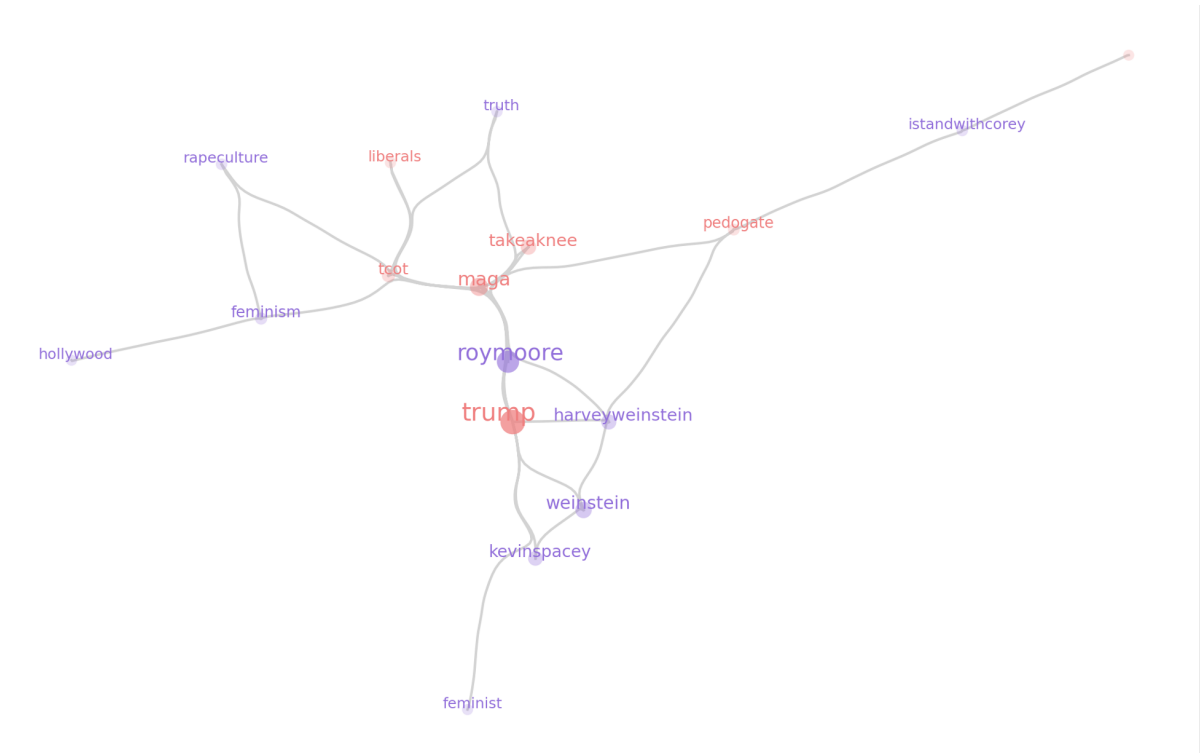


Fig. 4: The second largest hashtag community among tweet amplifiers.

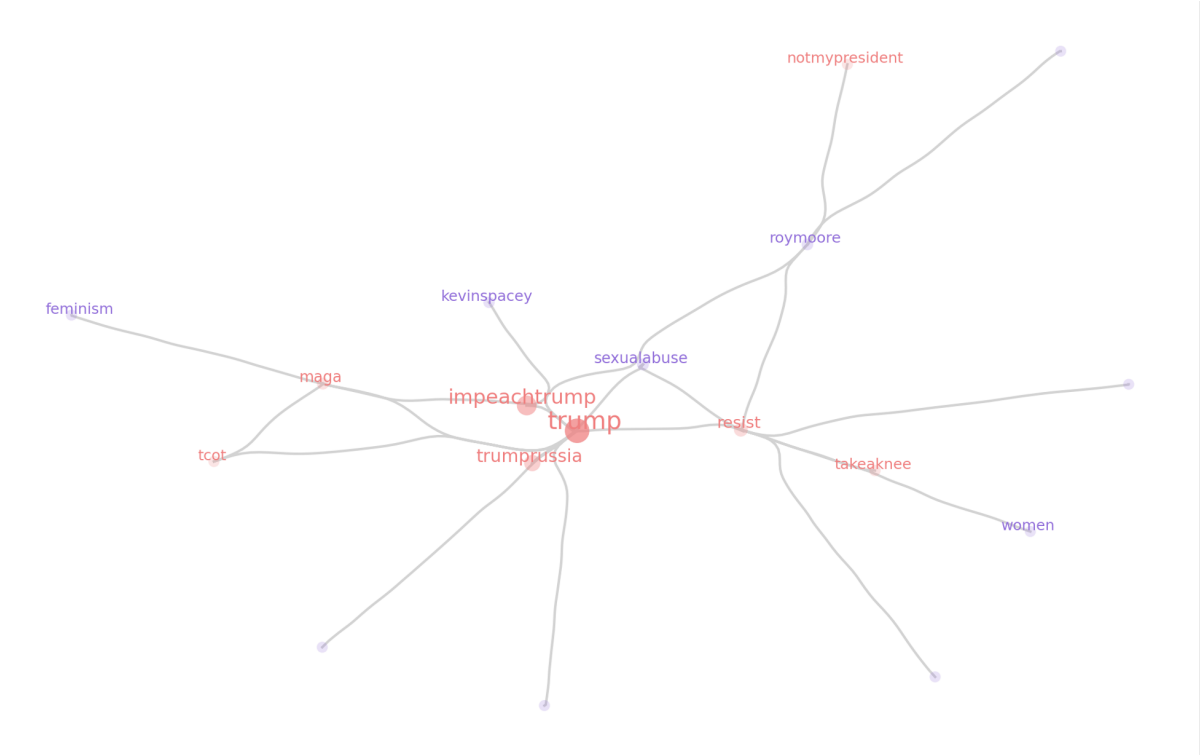


Fig. 5: The largest hashtag community among tweet contributors.

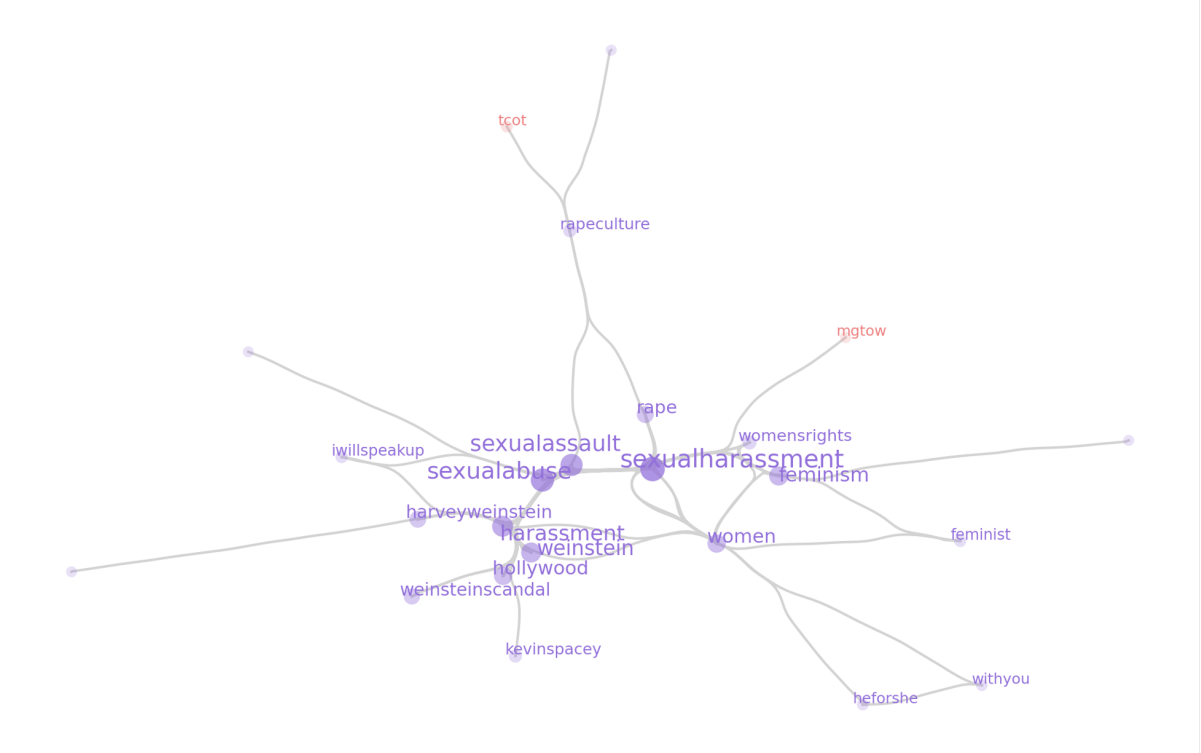


Fig. 6: The second largest hashtag community among tweet contributors.

exposure yet widening the scope of the discourse to the extent that the movement can become part of an ideological debate.

Our findings from the semantic network analyses provide further evidence to support the results of the OLS regression model and offers a more nuanced look into the underlying discursive patterns. The most notable outcome of the OLS model was the gradual decline in MeToo-related discourse among amplifiers, which contrasts with the more rapid increased association with politically relevant discourse. Our semantic network analysis offers a more structural insight into how these themes were organized and interconnected within the amplifier and contributor groups.

Within the contributor community, the network structure is characterized by a near-total saturation of MeToo-relevant hashtags. Central nodes such as #sexualassault, #harassment, and #womensrights form a dense and tightly clustered core, indicating that contributors sustained a coherent thematic focus aligned with the movement's original intent. The prominence and interconnectivity of these hashtags suggest that contributors were actively shaping and reinforcing a narrative centered on gender-based violence and the core issues of the #MeToo movement.

In contrast, the amplifier network reveals a markedly different structure. Here, general political hashtags, such as #trump, #impeachtrump, #resist, and #maga occupy the most central and interconnected positions in the graph. While MeToo-related hashtags are present, they appear more peripherally located and structurally dependent on the dominant political discourse. This suggests that amplifiers tended to position the #MeToo narrative within a broader political framework, often using it as a springboard to introduce ideological debates or partisan critique rather than as an issue of primary concern.

Taken together, these semantic differences reinforce our earlier finding that amplifiers and contributors played different roles within the broader discourse on the #MeToo movement. While contributors sustained a dedicated and thematically coherent engagement with the core issues of the #MeToo movement, amplifiers shifted the conversation toward broader political themes, effectively reframing the movement within the context of partisan politics. The results of our study offer strong empirical support that online social movements, though seemingly connected under the common #MeToo hashtag, are composed of diverse participant roles with divergent communicative goals and narrative trajectories.

Furthermore, our study underscores the importance of examining digital social movements not only through the lens of openness and diversity but also in terms of their inherent vulnerabilities to discursive control and re-appropriation. Social media offers a low-barrier entry point for a diverse array of users to engage in political discourse, which is the quintessential feature of digital activism. Nevertheless, the predominance of partisan hashtags among the contributors shows that the conversation around #MeToo can be shifted to reflect ideological intentions.

Practically, these insights offer valuable implications for strategic communication efforts by activists and advocacy organizations. Recognizing the susceptibility of movements to politicization, organizers may benefit from developing clearer narrative boundaries and proactive framing strategies to safeguard message coherence.

Moreover, our distinction between amplifiers and contributors offers a novel analytic framework for examining the dynamics of digital discourse. This differentiation captures two fundamentally distinct communicative strategies. Contributors engage in original content creation, often driven by motives such as opinion expression and emotional storytelling. Amplifiers, in contrast, facilitate the circulation of existing content, emphasizing the dissemination and visibility of the hashtags. Importantly, our classification operates at the level of the tweet, rather than the user. This method acknowledges that individual participants may assume multiple roles over time. By focusing on the communicative act rather than assuming a fixed user identity, we offer a more flexible and nuanced understanding of the participants in this discourse. While prior research has largely focused on aggregate patterns of engagement, our approach underscores the importance of communicative actions in tracing how social movement narratives evolve and transform over time.

6 Limitations and Future Directions

We recognize several limitations of this study. First, our classification of users as contributors or amplifiers is based solely on tweet content and surface-level tweet features, rather than user history or network behavior. Consequently, our analysis does not capture the broader influence of individual users as mediated through follower networks or relational structures. It was our intention to avoid focusing on offline or demographic identities of users in our dataset. This was partly a methodological choice. Twitter data often lacks reliable, verifiable indicators of user identity. In addition, our focus was on the roles enacted

through communicative practices rather than the attributes of the individuals behind them. A user can act as both a contributor and an amplifier at different points in a movement, which led us to classify roles at the "tweet-level" rather than the "user-level." Future research could integrate metadata or external information that can be validated to explore why some users more consistently contributed while others primarily amplified.

Second, our reliance on hashtag co-occurrence to infer framing strategies may not fully capture the nuances embedded in the tweet text. Hashtag patterns provide a useful approximation of discourse patterns, but they may overlook the affective tone or nuances that are present in the language itself. Future research could build on our approach by incorporating full-text analysis, sentiment detection, and user-level metadata to offer a more comprehensive view of how contributors and amplifiers shape digital discourse over time.

Third, while our analysis focused on the users' communicative roles (contributors vs. amplifiers), we acknowledge that Twitter's algorithmic ranking and recommendation systems could have played another important role in shaping visibility. Given that Twitter's algorithms are proprietary, we cannot directly measure these effects, but we anticipate that the platform infrastructure of any social media platform would naturally play a role in determining whose voices gained traction.

Fourth, we limited our scope to Twitter because it was at the time the most accessible and widely used platform for hashtag activism research. Twitter's public API enabled systematic data collection in a way that was not possible for Facebook, Instagram, or WhatsApp during the study period. Yet given that activists often coordinate messaging across multiple platforms, future studies that account for these multi-platform dynamics could better capture the nature of hashtag activism. We acknowledge that our dataset is from 2017 and that the social media environment has changed considerably. Twitter has re-branded as X and introduced new features and seen shifts in communities and moderation policies. While these shifts likely have impacted the mechanisms through which narratives are shaped on the platform, the findings remain valuable in understanding the development of digital activism practices for our chosen hashtag movement (#MeToo) and would serve as a baseline against which more recent cases can be compared.

Our findings serve as a timely reminder that the success of a digital movement is not solely determined by its virality or visibility. Rather, it is also contingent on who holds the power to define its meaning and how the narrative is either sustained or strategically subverted over time. By disentangling the distinct roles of contributors and amplifiers, our study takes an important step toward understanding the mechanisms through which digital narratives are constructed and contested, which ultimately impacts the perceived success of a social movement.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Bae SY, Kim T, Ha Yi, et al (2021) The medium and the backlash: The disparagement of the# metoo movement in online public discourse in south korea. International Journal of Communication 15:768–791

Bhatia S, Walasek L (2016) Event construal and temporal distance in natural language. Cognition 152:1–8

Blondel VD, Guillaume JL, Lambiotte R, et al (2008) Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10):P10008

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Clark R (2016) "hope in a hashtag": The discursive activism of #whyistayed. Feminist media studies 16(5):788–804

Entman RM (1993) Framing: Towards clarification of a fractured paradigm. Journal of Communication 43(4):51–58

Freelon D, McIlwain CD, Clark M (2016) Beyond the hashtags: #ferguson, #blacklivesmatter, and the online struggle for offline justice. SSRN (2747066)

Hurter C, Ersoy O, Telea A (2012) Graph bundling by kernel density estimation. Computer Graphics Forum 31(3pt1):865–874

Iyengar S (1990) Framing responsibility for political issues: The case of poverty. Political behavior 12(1):19–40

Jackson SJ, Bailey M, Welles BF (2020) #HashtagActivism: Networks of race and gender justice. Mit Press, Cambridge, Massachusetts

Katz E, Blumler JG, Gurevitch M (1973) Uses and gratifications research. Public Opinion Quarterly 37(4):509–523

Li M, Qin J, Jiang T, et al (2021) Dynamic relationship network analysis based on louvain algorithm for large-scale group decision making. International Journal of Computational Intelligence Systems 14(1):1242–1255

Liu J, McLeod DM (2019) Counter-framing effects of user comments. International Journal of Communication 13:2484–2503

Nguyen DQ, Vu T, Nguyen AT (2020) BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp 9–14

Nikolayenko O (2019) Framing and counter-framing a peace march in russia: The use of twitter during a hybrid war. Social Movement Studies 18(5):602–621

Rafieian O (2023) Optimizing user engagement through adaptive ad sequencing. Marketing Science 42(5):910–933

Smith AN, Seiler S, Aggarwal I (2023) Optimal price targeting. Marketing Science 42(3):476–499

Snefjella B, Kuperman V (2015) Concreteness and psychological distance in natural language use. Psychological science 26(9):1449–1460

Traag VA, Waltman L, Van Eck NJ (2019) From louvain to leiden: Guaranteeing well-connected communities. Scientific reports 9(1):1–12

Tufekci Z (2017) Twitter and tear gas: The power and fragility of networked protest. Yale University Press, New Haven, Connecticut

Introduction to the Track on Countering Information Manipulation Techniques

Joshua A. Tucker[1], Zeve Sanderson[1]

[1] Center for Social Media and Politics, New York University, New York, USA.

Corresponding authors. E-mails: joshua.tucker@nyu.edu; zeve.sanderson@nyu.edu

Abstract

This article introduces the thematic track on countering information manipulation techniques. It outlines the goals, approaches, and central questions addressed by the contributions within this track, which collectively explore strategies such as counter-speech, provenance and labeling initiatives, community-based fact-checking, regulatory interventions, and the implications of generative AI for information integrity.

Keywords

Information Manipulation, Extremism, Counter-Speech, Provenance and Labeling, Community-Based Fact-Checking, Generative AI, Regulation.

1 Introduction: The Challenge of Information Manipulation

The section on *Countering Information Manipulation Techniques* at the 2025 Paris Conference on AI and Digital Ethics was convened against the backdrop of intensifying concerns about the role of information in contemporary political and social life. The section was animated by the use of digital information—and digital information systems—as instruments of power: deployed to discredit political, ideological, and institutional opponents; to attempt to polarize populations; and even in some cases to undermine democratic trust. In such a context, practitioners, policymakers, and academics have aimed to understand how manipulation works and how it might be countered (Campbell and Gear, 1995).

The call for papers invited contributions that explored the mechanisms by which manipulated information spreads; the psychological dynamics that shape the actions of producers, spreaders, and believers; and the structural transformations of the information environment, including the rise of generative AI. Equally important, the track sought novel ideas for interventions, ranging from platform-based moderation frameworks to community review systems and new forms of accountability.

The four contributions in this section represent distinct but complementary approaches to these issues. Taken together, they offer insight into extremist exploitation of digital cultural spaces, the promises and pitfalls of grassroots fact-checking, the limits of counter-speech in the age of AI, and the regulatory innovations associated with China's AI content labeling regime. Collectively, they underscore the complexity of countering manipulation and the need for approaches that bring together studies of empirical rigor and normative questioning.

In the opening lecture at the start of this section of this conference, *Combating Online Harms: The Importance of Rigorous Scientific Testing of Proposed Interventions*, it was argued that interventions that appear promising in principle often disappoint in practice, and in some cases produce counterproductive effects.[4] The studies presented—ranging from experiments on online searching to evaluations of source credibility labels and algorithmic changes on Facebook and Instagram—reinforced the need for rigorous empirical evaluation before interventions are widely adopted. The papers in this section resonate with that argument: each provides insight into possible responses to information manipulation, but each also highlights the limits, trade-offs, and unintended consequences that accompany even well-intentioned solutions.[5]

2 Extremist Infiltration in Gaming and the Metaverse

Didier Danet's *Gaming the System: Extremist Infiltration Strategies in MMORPGs and the Metaverse* examines the ways in which extremist groups have embedded themselves within massively multiplayer online games and anticipates how these tactics may develop in emerging metaverse platforms. The paper asks how extremists adapt their strategies to exploit environments designed primarily for entertainment, and what lessons this holds for future digital spaces. Drawing on case studies and comparative analysis of online games and metaverse architectures, the paper details how extremists employ in-game chat, guild structures, custom avatars, and symbolic events to normalize radical narratives. These tactics, however, are constrained by developer control and player expectations.[6]

Looking forward to how this will play out in the metaverse, which the author does believe will play a growing role online in the future despite serious growing pains, Danet contrasts centralized metaverses, which offer access to large audiences but maintain moderation, with decentralized metaverses, which provide greater autonomy but limited reach, highlighting a dynamic that extremists must navigate when developing propaganda strategies.

The author suggests that extremists will likely continue to infiltrate mainstream platforms rather than confining themselves to marginal spaces, despite the advantages of the latter in terms of lack of content moderation. This has important implications for the detection and mitigation of extremist influence in increasingly immersive environments. By situating manipulation within gaming culture and metaverse

[4] Tucker gave the opening lecture of this section of the conference, while Sanderson oversaw the process of soliciting reviews for the papers.

[5] In the spirit of the conference, we composed this introductory section with the assistance of Chat GPT 5 and want to be completely transparent about having done so. The suggested text from ChatGPT was carefully reviewed and heavily edited/rewritten both to ensure that it reflects our views and that it is factually accurate. We did so both for reasons of efficiency – despite heavy editing the process was likely faster than if we had not used ChatGPT – and understanding, as the conference sparked in us a desire to see how AI could be used for such a task and where problems could arise in doing so. The process was illuminative for both of us, to say the least.

[6] For references to relevant literature, please see the paper itself.

design, the paper demonstrates how propaganda adapts to shifting cultural norms and technological affordances.

3 Community-Based Fact-Checking as Socioinformatic Intervention

A Socioinformatic Analysis of Community-Based Fact-Checking—co-authored by Julian Pelloth, Lena Pölzer, and Katharina Zweig—addresses the question of whether Community Notes can effectively counter misinformation on X (formerly Twitter), contributing to a growing literature on the promises and limitations of crowdsourced efforts for information integrity (Godel, Sanderson, Aslett, Nagler, Bonneau, Persily, and Tucker, 2021; Martel, Allen, Pennycook, and Rand, 2024; Pennycook and Rand, 2019; Slaughter, Peytavin, Ugander, and Saveski, 2025). The authors formulate a research question centered on the fact that the vast majority of notes fail to achieve a definitive status of either "helpful" or "not helpful". Their analysis of publicly available data reveals that only a small fraction of notes are ever displayed given that the majority remain in the intermediate status of "needs more ratings." Though this system may be by design,[7] the authors—after a thorough explanation of how Community Notes' rating system works—suggest that this reflects a fundamental ineffectiveness in the system's ability to moderate content and prevent the spread of misinformation as intended.

To diagnose the cause of this ineffectiveness, the authors employ a socioinformatic analysis, a method designed to model the interactions between social actors, their motivations, and the technical system itself. The analysis identifies the key social actors as the platform, X, which aims to reduce misleading posts, and the contributors who rate the notes. The authors hypothesize that many raters may be motivated not by a note's objective helpfulness, but by whether the note aligns with their pre-existing beliefs.

This dynamic creates a vulnerability to manipulation by suppressing factual, corrective notes. By creating artificial dissent through coordinated "not helpful" ratings, manipulative actors can prevent a note from ever achieving the "helpful" status required for it to be seen by the public, thus ensuring the original misleading post remains unchallenged.

4 The Limits of Counter-Speech in the Age of AI

Isaac Taylor's *Speech Act Gaps and Counter-Speech* approaches information manipulation from a philosophical perspective, asking whether traditional liberal remedies remain effective in an era of AI-generated communication. Taylor's central claim is that artificial speech often produces a "speech act gap": utterances that resemble speech acts but cannot be attributed to an intentional actor.[8]

Taylor argues that this phenomenon can undermine the efficacy of counter-speech. Correcting or sanctioning a human speaker can deter future harmful speech by showing the speaker that there may be public consequences from the speech action. But when dangerous speech is produced by a language model, counter-speech can lose these benefits. The initial utterance may still cause harm—by spreading stereotypes, reinforcing biases, or disseminating misinformation—but counter-speech may not operate as intended because the deterrent effect of counter-speech on humans who are concerned about public sanctions may not hold for AI-generated speech.

Taylor concludes that reliance on counter-speech may therefore falter in the age of synthetic communication, and that direct restrictions on AI outputs are worth considering, especially as they may not raise the same free speech concerns that attach to human expression. For our section, this contribution highlights new dynamics that may arise when machines are employed rather than humans to generate harmful or manipulative content, and it underscores the need to rethink normative frameworks that were designed for a different communicative environment.

5 Lessons from China's AI Content Labeling Mandate

Wenlong Li's *Accountability and Attribution in AI-Generated Content Authentication: Lessons from China's AI Content Labelling Mandate* analyzes China's recently implemented regulatory framework for AI content provenance. Effective in September 2025, this regime mandates both visible and metadata-based labeling of AI-generated content, embedding compliance obligations throughout networks for AI

[7] See https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes.
[8] For references to relevant literature, please see the paper itself.

content supply and distribution. A growing body of research across law, computer science, and the social and behavioral sciences is concerned with content labeling (Chesney, 2023; Gils, 2024; Meding and Sorge, 2025; Simon, Fletcher, and Nielsen, 2024), and this paper provides a timely analysis of this developing policy area.

The analysis situates China's framework within the broader international regulatory landscape, contrasting it with the European Union's more context-sensitive AI Act and the fragmented approaches found in the United States. While the Chinese model represents "an unprecedented statutory response to the complex, polycentric reality of AI-generated content to operationalize accountability at scale" (p. 3), it is not without vulnerabilities. Metadata can be stripped or falsified, interoperability across jurisdictions is limited, and rigid mandates risk chilling creative expression. Nonetheless, the Chinese approach demonstrates one path toward addressing the "many hands" problem of AI accountability, which as it is rolled out will offer an important example for regulators worldwide who are designing policies to combat AI manipulation.

6 Cross-Cutting Themes

Although these four contributions differ in method—ranging from computational analysis to normative philosophy and policy evaluation—they converge on several important themes.

Adaptation of manipulation. Each paper demonstrates that manipulation exploits cultural contexts and technological affordances. Extremists leverage the social dynamics of gaming, community fact-checking contends with the affordances and individual incentives of participatory models, AI systems introduce novel forms of speech without speakers, and regulators attempt to inform the public when "to see is to believe" has been undermined. Manipulation is not static; it continually adapts to new technical architectures and social practices.

Centering trust in the study of information manipulation. Across the contributions, trust emerges as a central battleground. Whereas a decade of research has largely focused on distinguishing between true and false as an objective measure, these papers contribute to a renewed focus on trust as the foundation to the information environment. What vectors of manipulation can be exploited? Who do we trust to protect us against harmful speech? And what can we do to provide the public with the necessary information to trust content they see?

Limits and trade-offs of interventions. Each paper underscores that interventions are imperfect and often entail unintended consequences. Community fact-checking may be hijacked; counter-speech may lose efficacy against synthetic utterances; provenance mandates may be circumvented or suppress legitimate expression; and gaming moderation may fail to prevent extremist exploitation. This aligns with the central argument of Tucker's keynote: interventions must be tested rigorously, and we must resist the allure of simple solutions to complex problems.

The themes of this section resonate directly with the arguments Tucker advanced in his introductory section address, as he stressed that the field must move beyond general enthusiasm for interventions to rigorous empirical testing of their efficacy. The papers collected here reinforce that imperative. By bringing these diverse perspectives into conversation, this section aims to advance the scholarly and policy dialogue on countering information manipulation techniques. It is our hope that the analyses presented here will contribute to the development of more effective, more accountable, and more ethically grounded strategies for addressing one of the defining challenges of the digital age.

Acknowledgements

We thank Hubert Etienne for inviting us to be part of PCAIDE, Thomas Souverain for his support throughout the process, and many anonymous reviewers for their assistance in selecting papers for the section.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Campbell SL, Gear CW (1995) The index of general nonlinear DAES. Numer Math 72(2):173–196

Chesney R (2023) Should the united states or the european union follow china's lead and require watermarks for generative ai? Georgetown Journal of International Affairs URL https://gjia.georgetown.edu/2023/05/24/should-the-united-states-or-the-european-union-follow-chinas-lead-and-require-watermarks-for-generative-ai/

Gils T (2024) A detailed analysis of article 50 of the eu's artificial intelligence act. Chapter to appear in an upcoming commentary on the EU AI Act (Q3-4 2024)

Godel W, Sanderson Z, Aslett K, et al (2021) Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. Journal of Online Trust and Safety 1(1)

Martel C, Allen J, Pennycook G, et al (2024) Crowds can effectively identify misinformation at scale. Perspectives on Psychological Science 19(2):477–488

Meding K, Sorge C (2025) What constitutes a deep fake? the blurry line between legitimate processing and manipulation under the eu ai act. In: Proceedings of the 2025 Symposium on Computer Science and Law, pp 152–159

Pennycook G, Rand DG (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences 116(7):2521–2526

Simon F, Fletcher R, Nielsen RK (2024) What does the public in six countries think of generative ai in news? URL https://reutersinstitute.politics.ox.ac.uk/what-does-public-six-countries-think-generative-ai-news

Slaughter I, Peytavin A, Ugander J, et al (2025) Community notes reduce engagement with and diffusion of false information online. Proceedings of the National Academy of Sciences 122(38):e2503413122

Gaming the System: Extremist Infiltration Strategies in MMORPGs and the Metaverse

Didier Danet[1,2]

[1] GEODE Science Center, Paris 8 University, Saint-Denis, France.

Corresponding author. E-mail: didier.danet@gmail.com

Abstract

Extremist groups strategically infiltrate digital spaces where large audiences congregate, adapting their propaganda tactics to exploit platform-specific characteristics. This parasitic approach proves particularly effective within massively multiplayer online games (MMORPGs), where extremists weaponise game mechanics—chat systems, character customisation, and community features—to cultivate receptive audiences whilst navigating developer-imposed constraints.

These operations encounter a fundamental tension: the need to soften propaganda messages to evade platform bans conflicts with extremists' desire for explicit indoctrination and players' entertainment-focused expectations.

Metaverses represent the next propaganda battlefield, offering unprecedented opportunities that surpass traditional MMORPGs. Enhanced immersivity, simplified content creation, enriched non-verbal communication, and sophisticated emotional regulation capabilities will significantly amplify propagandist potential. Centralised metaverses will likely mirror existing gaming environments, whilst decentralised platforms promise extremists greater freedom to construct radicalised virtual worlds.

Yet the audience autonomy dilemma persists: decentralised metaverses offer ideological control but struggle to attract mainstream users who gravitate toward polished, entertainment-focused centralised platforms. This tension between mass appeal and radical control will continue shaping extremist strategies, forcing groups to balance unrestricted propaganda dissemination against accessing substantial audiences.

Understanding these dynamics proves crucial for developing effective counter-strategies as digital landscapes evolve toward increasingly immersive and manipulable environments.

Keywords

Metaverse, MMORPG, Extremist Groups, Follow the Audience Strategy.

1 Introduction

When Mark Zuckerberg launched the Metaverse project in 2021, positioning it as the next digital revolution, few anticipated the spectacular failure that would follow (Tassi, 2025). Despite a staggering 20 billion investment and the symbolic rebranding of Facebook to Meta, Horizon Worlds now attracts fewer than forty daily users. This dramatic collapse can be attributed to a confluence of factors: technological limitations (cumbersome headsets, mediocre graphics), lack of concrete applications, absence of significant added value for users, and competition from generative AI that diverted attention elsewhere. With annual losses exceeding 13 billion and no foreseeable return on investment, Meta's flagship venture serves as a cautionary tale of premature deployment and misguided vision.

However, the failure of this initial metaverse implementation does not condemn the concept itself. Meta's misstep merely demonstrated that a vision lacking originality and deployed prematurely was not viable. Rather than pursuing a monolithic "all-in-one" platform, the industry now favors specialized applications (gaming, virtual fashion, professional training) and concrete use cases. As the underlying technologies mature and applications emerge that genuinely address user needs, the Metaverse is poised to become the new paradigm of digital space in the 21st century (Benaben, Congès, and Fertier, 2025).

An online metaverse provides access to real-time, shared, and persistent 3D space simulations where users can collectively experience immersive environments (Basdevant, François, and Ronfard, 2022). Compared to our current digital landscape, Metaverse development promises several compelling advantages: enhanced immersion that strengthens user engagement through 3D environments that create a sense of physical "presence"; enriched social interactions through avatars and virtual spaces that enable more spontaneous and expressive communication than current videoconferencing; spatial collaboration that makes working together in shared 3D environments more intuitive than existing tools, and unified experiences that create seamless continuity between different applications and platforms.

Yet these undeniable advantages come with a critical counterpart: the Metaverse will present unprecedented opportunities for entities conducting influence operations and disinformation campaigns (Pira, 2024; Wang, Makowski, Cieślik, Lv, and Lv, 2023). The blending of reality and fiction, coupled with the emotional and sensory manipulation possible in virtual environments, will endow influence operations with particularly powerful persuasive force. Virtual reality universes will enable psychological and emotional manipulation of users at levels far exceeding what we experience today, facilitating identity theft of public figures, amplification of disinformation through automation and multiple account creation, and targeted manipulation using realistic digital humans capable of establishing fraudulent relationships with digital audiences. It is therefore imperative to anticipate the risk of influence operations in the Metaverse by examining strategies that could exploit its distinctive characteristics. Although instantiations of the Metaverse concept remain largely in development—particularly following Horizon Worlds' failure—their main contours can be discerned by examining online games (Massively Multiplayer Online Role-Playing Games), which represent the most advanced prefigurations of this concept. Despite their ludic purpose and architectural differences from conventional social networks, these games serve as privileged terrain for extremist groups that deploy specific strategies to promote their doctrines and ideas among young players. We can reasonably assume that these groups will adapt their influence strategies in the Metaverse by exploiting the technical possibilities afforded by advancing technology.

The first part of this analysis describes the strategies implemented by extremist groups to infiltrate online gaming environments and deploy their influence operations. This question has been the subject of numerous quantitative and qualitative studies that provide convergent and sufficiently robust results, making it unnecessary, within the scope of this article, to conduct a specific investigation. Therefore, the presentation of strategies employed by extremist movements will rely on the existing literature, especially the recent works by Allchorn and Orofino, (Allchorn and Orofino, 2025) Davey, (Davey, 2024) as well as Schlegel and Robinson. (Schlegel and Kowert, 2024a)The second part endeavors to anticipate, based on evidence presented in the first section, what adaptations of these strategies might emerge in future Metaverse instantiations.

2 Quest for Influence: How Extremists Target and Adapt to Gaming Audiences

The active presence of far-right political movements in the world of multiplayer online games is all but a surprise. Extremist groups have seized upon these spaces to create "radicalization pipelines", as they have long done on YouTube and other social media platforms. (Champion and Frank, 2021; Ledwich and

Zaitsev, 2019; Mamié, Horta Ribeiro, and West, 2021; Ribeiro, Ottoni, West, Almeida, and Meira Jr, 2020) Although the existence of radicalization pipelines is less well known within the MMORPGs' universe, it has been the subject of research and reports by academics and citizens groups that have clearly highlighted its significance and characteristics. (Condis, 2019,2; Wells, Romhanyi, Reitman, Gardner, Squire, and Steinkuehler, 2024). It is part of a broader trend that sees a new generation of activists chasing existing audience especially in the field of popular culture events, gatherings or conventions : car tuning (Darras, 2020) , MMA (Nissen, Avramov, and Roberts, 2021; Vysotsky, 2022) or even marriage coaching... All these forms of popular culture are ignored or even despised by major cultural institutions but they bring together groups presenting personal, social, and cultural characteristics likely to resonate with extremist themes and which are as many potential recruitment pools.

This generic infiltration strategy has therefore been adapted to the universe of online games where it has found favourable ground, even though it encounters obstacles related to platform architecture and the awareness of their designers.

2.1 Gaming the System: How MMORPGs Environments Enable Propagandist Tactics

Online games present several characteristics that make them privileged spaces for influence operations (Silva, Valadares, Pedrosa, Rezende, Cappelle, and Assis, 2023; Williams, Kirschner, and Deterding, 2024) .

The population of online gamers has long been stereotyped as young men or adolescents confined to their rooms, playing war-themed games such as Call of Duty or World War II titles. However, recent studies show that this image has evolved considerably. The global number of gamers now approaches 3.5 billion (Andrea, 2025), far exceeding the adolescent age group. In fact, the average age of players is 36 years, with a significant presence of gamers aged over 50. The proportion of female players has increased dramatically, particularly within so-called "casual" games commonly played on mobile devices such as smartphones.

However, despite these demographic changes, the segment of young men, who constitute the primary target audience for extremist movements, remains highly attractive. For example, 70 percent of young American men are estimated to regularly engage in online gaming. These individuals spend multiple hours per week playing and tend to remain loyal to the same game for several years. Their sustained and intensive engagement represents a significant advantage for extremist groups, which can deploy long-term influence and recruitment operations and progressively modulate the radicalization process through a gradual, non-confrontational approach that avoids aggressive pressure on individuals.

Secondly, gaming culture often tolerates—sometimes even celebrates—"edgy" humor, competitive aggression, and transgressive expressions, including sexism, racism, homophobia, and "politically incorrect" language. This ambient toxic environment enables actors with radical beliefs to blend in, test boundaries, and progressively advance extremist narratives that would be unacceptable on more strongly moderated platforms. (Boudreau, 2022; Donner, 2024; Munn, 2023). It is important to emphasise that the online gaming community cannot be regarded as uniformly aligned with extremist, masculinist, or racist ideologies in its entirety. Nevertheless, "gaming culture" exhibits certain specificities that create exploitable entry points for radical propaganda. Both players and game developers often engage with a culture that valorises transgression, symbolic violence, provocative humour, and the use of ambiguous or coded references. Online forums associated with gaming platforms, such as Steam and Reddit, have witnessed the emergence of radicalised sub-communities; a notable example being the r/KotakuInAction forum, which has connections to the Gamergate movement (Ferguson and Glasgow, 2021). Moreover, the frequent deployment of memes and jokes with racist or sexist undertones in games like Grand Theft Auto Online (Plans Vilaseca, 2024) or The League of Legends (Aguerri, Santisteban, and Miró-Llinares, 2023) often acts as an footstep to more explicit radical discourse promoting violence (Sorell and Kelsall, 2025). These cultural dynamics illustrate how seemingly innocuous interactions within gaming spaces can serve as vectors for the gradual normalisation of extremist viewpoints.

A key catalytic factor is the relative weakness of moderation. Many gaming and gaming-adjacent platforms lack robust systems to monitor and filter content, giving space for hate speech, propaganda, and radicalization to go unchecked for extended periods. (Kilmer, Aslan, and Kowert, 2024) Unlike mainstream social networks, gaming platforms typically prioritize user autonomy and game functionality, often at the cost of effective content regulation. This allows organic extremism to flourish, as individuals exploit loopholes in moderation and the anonymity afforded by these environments.

In manifestation, organic extremism on gaming platforms takes several forms: 1. Users sharing hateful memes or jokes in chats or forums. 2. Participation in harassment or doxxing campaigns targeting minorities or perceived opponents. 3. Informal formation of communities with shared radical beliefs, where extremist discourse becomes normalized and reinforced. 4. The adaptation of gaming experiences—such as in-game symbols, avatars, and narratives—to signal affiliation with extremist ideologies and attract others.

In opposition to organic extremism which emerges spontaneously from individuals or loosely connected peer groups within gaming spaces, organizational extremism involves deliberate, coordinated actions by established extremist groups or networks that strategically exploit gaming environments to advance their agendas. Extremist organizations use six main identified strategies to capitalize on the unique characteristics of large gaming platforms to further their agenda. 1. Gamification : Organizational actors strategically incorporate game mechanics such as points, leaderboards, badges, and "achievements" into their digital propaganda tools. (Lakhani and Wiedlitzka, 2023) These gamified elements incentivize engagement, encourage participation, promote competition among supporters, and thereby foster deeper identification with extremist content. For example, extremist groups develop apps or platforms rewarding "missions" or tasks aligned with their objectives. 2. Production of Bespoke Video Games : Extremists produce their own tailor-made games delivering immersive experiences that embed ideological narratives. These games aim to increase attraction and familiarity with extremist ideas, presenting them in an entertaining format that lowers resistance and facilitates recruitment. For instance, far-right developers have created games where players defend ideological positions or simulate battles aligned with extremist worldviews. 3. Modification of Existing Games : Rather than creating games from scratch, extremists often develop modifications (mods) for popular existing titles to insert extremist symbolism, narratives, or scenarios. (Robinson and Whittaker, 2020) Mods allow leveraging large player bases and the familiarity of established game worlds to subtly or overtly introduce radical content, as seen in examples like maps simulating terrorist attacks or creation of white ethnostates in widely played games. (Schlegel, Wiegold, Winkler, and Junk, 2025) 4. Use of In-Game Chats : Organized extremist groups exploit in-game communication channels by creating private or semi-private chat rooms within games to coordinate, share propaganda, recruit, and groom potential followers. These chats allow for discreet, trust-based community-building that is difficult for moderators to monitor effectively. Unlike organic individual users, organizational actors systematically use these digital spaces as communication infrastructures integral to their strategic operations. 5. Presence on Gaming-Adjacent Platforms : Beyond the games themselves, extremist groups maintain coordinated presences on platforms linked to gaming culture, such as Discord, Twitch, Steam, Reddit, and chan boards. (Kowert, Kilmer, and Newhouse, 2024) These platforms provide tools for mass communication, live streaming, organization of real-world events, and persistent propaganda dissemination. Organizational actors exploit the technical affordances and often lax moderation policies of these platforms to form echo chambers, mobilize supporters, and escalate recruitment. 6. Gaming Cultural References : Extremists adopt gaming aesthetics, imagery, language, and footage in their propaganda outputs to resonate culturally with target audiences. (Munn, 2019) This strategic use of familiar gaming symbols and tropes aids in attracting and normalizing extremist ideas within receptive communities without immediate detection.

This presentation allows for a precise delimitation of the scope of this study, which is limited to the analysis of the organizational strategies implemented by extremist movements within the realm of online video games, with a particular focus on the combined use of communication channels present in the games themselves (in-game chats) and adjacent platforms to which interested players are directed.

2.2 Developer Control vs. Propagandist Ambition: Structural Constraints in MMORPGs

Despite this fertile ground for radicalization, the expansion of extremist activities within these digital ecosystems increasingly encounters resistance from their creators (League, 2024). Platform operators have become more aware of the threats posed by extremist exploitation, and a growing number of them have taken decisive measures to mitigate these risks. (Schlegel, 2021) This includes enhancing content moderation capabilities, deploying artificial intelligence to detect extremist content, banning extremist-associated accounts, and collaborating with governmental and non-governmental organizations dedicated to counter-extremism. Nonetheless, the challenge remains significant due to the size and complexity of gaming environments (Dendi, 2025) (Gorwa, Binns, and Katzenbach, 2020) and the adaptability of extremist actors who continually refine their tactics to evade detection.

The strategy of infiltrating an audience assumes engaging with games whose propagandists are not the original creators, as initial attempts at direct control failed rapidly. Even if gamers are given large customization opportunities (Adinolf and Turkay, 2011), the developers of the infiltrated games maintain authority over the narrative, structure, and rules, and they tightly constrain the extent to which players can customize the gaming experience. (Sampedro, Grassi, Zitkus, and Valente, 2023) Major platforms actively ensure that these customization options are not exploited by extremist groups for propaganda or to divert players from the game's core objectives.

Although moderation efforts were initially very limited, platform managers have become aware of the existence of propaganda operations and the reputational damage risks they carry. (Lamphere-Englund, White et al., 2023) The growing concentration in the video game industry would allow a smaller number of powerful companies to exert greater control over their platforms. This would facilitate a more uniform and effective implementation of anti-extremism measures, as these companies would have ample resources to invest in advanced content moderation, research, and collaboration with governments and NGOs. Standardized policies across major platforms could reduce safe havens for extremist content, while increased public scrutiny might motivate companies to take greater responsibility for user safety. Additionally, industry concentration would simplify coordinated efforts to counter violent extremism. Overall, this concentration would enhance the sector's capacity to develop and enforce proactive, well-resourced strategies against extremist exploitation of gaming spaces.

Moreover, games in progress in the main online platforms are public and interactions between team members are easily accessible to other players. Some are even commented on live on specialised YouTube channels. This publicity increases the risk for propagandists who are not subtle enough to be spotted and reported to moderators.

Finally, as has been recalled, users of online gaming platforms have the dominant motivation of gaming pleasure; they do not participate in MMORPGs for ideological purposes. Launching a sexist joke, which is classically part of "Gaming culture", does not necessarily mean that its author is ready for deep engagement in extremist discourse and practice.

2.3 Monitoring the Virtual Battlefield: Detection Strategies for In-Game Propaganda

Building on initial approaches focused on content analysis, (Mattheis and Kingdon, 2023) a range of monitoring strategies can be employed to better understand, detect and address the risks associated with extremist activity in online games. One such strategy involves the systematic analysis of in-game chat conversations to detect keywords, linguistic patterns, or thematic content that may incite hatred or promote extremist ideologies. (Davey, 2024)These messages are often embedded within seemingly benign exchanges — such as strategic communication or casual interaction — enabling actors to disseminate radical discourse covertly under the guise of team coordination.

Methods	**Limitations**
Automatic detection of keywords and expressions	Rapid evolution of coded language ; specificity of games'vocabulary (ennemy, kill, eliminate)
Contextual analysis to detect codes and euphemisms	Risk of false positives with gaming humour
Identification of radicalisation language patterns	Difficulty analysing multilingual conversations
Monitoring linguistic variations circumventing filters	Large volume of data to process

Table 1: Analysis of In-Game Chat Conversations

In parallel, the monitoring of visual symbols — such as skins, flags, insignia, and other customisable elements — can provide critical insight into ideological affiliations. These visual markers, which are frequently integrated into avatars, in-game objects, or virtual environments, may signal alignment with banned or anti-democratic ideologies. (Steinkuehler and Squire, 2024) As an extension of this approach, it would be possible to establish a systematic surveillance framework to detect the presence and circulation of such symbols, especially within user-generated content and community-driven customisation features.

Methods	Limitations
Surveillance of character/guild names	Constantly renewed symbols
Identification of hijacked emblems and logos	Possible double meaning of references
Detection of coded numerical references	Difficulty distinguishing innocent use from hijacking
Analysis of avatars and customisations	Cultural variations in interpretation

Table 2: Monitoring of Visual Symbols

Another area of concern involves the formation and functioning of in-game social structures, such as teams, guilds, and clans. (Schlegel and Kowert, 2024b) Monitoring these organisational structures can help identify suspicious behavioural patterns, including the creation of insular or hierarchically structured groups that deviate from conventional player dynamics. These groups may serve as echo chambers for ideological reinforcement or as recruitment spaces deliberately crafted to foster a sense of belonging and loyalty to a radical cause. Analysis of group naming conventions, leadership hierarchies, and interaction frequency can offer valuable indicators of such activity.

Methods	Limitations
Analysis of suspicious recruitment patterns	Difficulty distinguishing legitimate from problematic recruitment
Identification of unusual hierarchical structures	Rapidly evolving recruitment techniques
Monitoring post-recruitment activities	Complexity of guild social structures

Table 3: Detection of In-Game Toxic Structures

Finally, the identification of manipulation patterns — particularly in the form of ideological grooming — represents a critical component of early detection. (Koehler, Fiebig, and Jugl, 2023) In the gaming context, grooming often involves a gradual process of trust-building and normalisation of extremist ideas, rather than overt proselytising. Techniques may include sustained one-on-one interactions, the use of emotionally resonant narratives, and the strategic deployment of in-game incentives to guide players toward radical viewpoints. Detecting these processes requires not only behavioural analysis but also a deep understanding of how social influence and ideological persuasion can be subtly embedded within gaming culture and interactions.

Methods	Limitations
Detection of ideological "grooming" techniques	Subtlety of manipulation techniques ; Cross-platform processes create discontinuities in content moderation
Detection of manipulative conversation patterns	Difficulty automating detection
Identification of attempts to normalise extreme ideas	Risk of confusion with normal conversations
Analysis of psychological destabilisation techniques	Constant evolution of tactics

Table 4: Identification of manipulation patterns

The methods for detecting infiltration strategies conducted by extremist groups in the universe of online games are therefore multiple. Some require limited efforts, such as analysis of explicit interactions (conversations in game chats). Others are much more ambitious and complex, such as identification of manipulation and indoctrination patterns. They suggest that a policy to combat these toxic practices is possible and could be much more vigorous than it currently is, even though it presents challenges of all sorts: economic, technical, human, and ethical.

Technical	Human	Ethical
Need for sophisticated AI systems	Moderator training	Privacy protection
Necessity of constant updates	Risk of team burnout	Risk of excessive surveillance
Management of large quantities of data	Need for extremism expertise	Questions of consent
Interoperability between different games	Linguistic and cultural barriers	Balance between security and freedom of expression

Table 5: Detection Challenges

3 Virtual Reality, Real Threats: Extremist Adaptation in the Metaverse Era

3.1 Centralised vs Decentralised Metaverse Models

The concept of the Metaverse can be adapted according to very different instantiations depending on the techniques implemented and their arrangement (De Rossi, Abbatemarco, Diaferia, and Salviotti, 2024; Mystakidis, 2022; Schöbel, Karatas, Tingelhoff, and Leimeister, 2023). Two main architectures can be distinguished: centralised and decentralised metaverses. The two types of metaverses represent two fundamentally different approaches to creating shared virtual worlds.

In a centralised metaverse, a single entity (generally a company with significant financial, technical, human resources...) creates and controls the infrastructure, servers, data, and rules (Meta, Roblox, or Fortnite Creative) (Ghosh, Hassija, Chamola, El Saddik et al., 2024; Karaarslan and Yazici Yilmaz, 2023). These metaverses are the best known and they are intended to gather the largest audiences. The choice of centralisation guarantees the company control of the metaverse it creates through a certain number of specific technical devices such as internal hosting of proprietary servers, centralisation of databases hosting user profiles and assets, closed APIs controlled by the company, or the implementation of traditional payment systems. Furthermore, the companies operating these platforms establish their terms of use and enforce codes of conduct backed by sanctions that can culminate in permanent account termination. These codes of conduct increasingly embed ethical behavioural standards and proscribe discourse that undermines democratic principles.

An illustrative but contested example is Twitch's removal of the livestream by the perpetrator of the Buffalo massacre in less than two minutes. On May 14, 2022, Payton Gendron, an 18-year-old white supremacist, livestreamed as he conducted a racially motivated shooting in a supermarket in Buffalo, New York, killing ten people. Despite Twitch's rapid action to take down the livestream within about 100 seconds, the incident exposed the persistent challenge of moderating live violent content online. (Grayson, 2022; Hern, 2022) This speed was a significant improvement compared to similar past incidents, such as the Christchurch mosque shootings in 2019, where it took 17 minutes for Facebook to remove the livestream. However, despite efforts of major platforms, preventing the spread of violent videos remains difficult given that clips are quickly edited, downloaded, and reposted across multiple platforms.

Regulation of centralized Metaverse platforms could then be strengthened through self-regulation driven by reputational risks and ethical concerns, coupled with external, legislative regulation similar to the requirements imposed on major platforms by the Digital Services Act regarding the moderation of extremist, hateful, or harassing speech These centralised systems would be therefore very similar in many respects to existing MMORPGs.

In decentralised metaverses , on the other hand, no single entity controls the whole structure (Decentraland, The Sandbox, or Somnium Space). The technological building blocks implemented are significantly different, such as the central place accorded to blockchain and non-fungible tokens to represent and manage user assets, decentralised protocols for data storage, open and interoperable standards...

The two main metaverse models will therefore offer a very different experience to their users. In centralised metaverses, users will benefit from greater ease of access and use as well as more fluid and optimised use provided by the platform manager. They will benefit from centralised technical support as needed. Online activities will be monitored and, where appropriate, they may be subject to more or less strict moderation depending on the company's policy.

In decentralised metaverses, users will participate in collective control of the organisation (Decentralised Autonomous Organisations) governed by a transparent system of rules and transactions.

Decisions taken will be slower and the truly collective character will undoubtedly be undermined by an abstention phenomenon as in any collective organisation of this type. Moderation will most often be absent or impracticable because it too is governed by a collective decision-making process.

Centralised metaverses facilitate user access to pre-established and optimised services but limit their autonomy because the company retains control of the structure and its evolution, whilst decentralised ones favour user autonomy who must accept the disadvantages of complex collective governance.

3.2 Unleashing Digital Deception: How Metaverse Features Enable Greater Manipulation

Whether centralised or decentralised, the different instantiations of the Metaverse will be privileged terrain for influence operations due to the new opportunities they carry. Compared to the universe of online video games that exist today, metaverses will indeed offer new capabilities. The increasing immersiveness of metaverses introduces new psychological dynamics that may be exploited by extremist propagandists. The concept of "body ownership" — the psychological sensation that a virtual avatar constitutes an extension of one's own body — is particularly important in this context (Ehrsson, 2020; Kilteni, Maselli, Kording, and Slater, 2015; Lam, Yang, Alhilal, Lee, Tyson, and Hui, 2022). Advanced synchronisation of movements and enhanced sensory feedback, such as virtual reality and haptic technologies, intensify the players' connection with their avatars (Kim, Lee, and Chung, 2023; Lee, Chang, Uhm, and Owiro, 2023). This heightened sense of embodiment potentially amplifies the impact of influence operations. For example, in decentralised metaverses, immersive role-playing scenarios could be designed to immerse users in extremist narratives, positioning them in situations that valorise violence against targeted groups, thereby normalising radical behaviour through experiential engagement.

Concurrently, advancements in affective computing (Coutinho and Boukerche, 2023; Pervez, Shoukat, Usama, Sandhu, Latif, and Qadir, 2024) enable more precise monitoring and interpretation of users' emotional states through technologies such as facial recognition, eye-tracking, and pulse measurement. The proliferation of increasingly sophisticated yet affordable connected devices facilitates this trend. These innovations allow content delivery to be dynamically adapted to the user's emotional responses — modifying the nature, style, or pacing of information presented — to optimise the progression along an indoctrination pathway. Such personalised modulation of content represents a novel vector for subtle ideological influence.

Moreover, the enrichment of social interactions within these digital environments — through gestures, spatial proximity, and personalised avatars — serves to strengthen users' sense of community and belonging. Subtle visual markers embedded in avatars, such as insignia or symbols, may function as discreet identifiers that enable the recognition and recruitment of like-minded individuals, further embedding ideological affiliations within social dynamics.

However, effective moderation remains a significant challenge in three-dimensional and immersive spaces. Traditional moderation tools, designed primarily for text and image-based platforms, are ill-suited to the detection of hate symbols and the analysis of real-time interactions in dynamic virtual environments. This technological gap allows extremist groups to exploit personalisation algorithms, which can entrench users within ideological echo chambers where radical content is disproportionately represented. The combination of algorithmic reinforcement and increased immersion exacerbates the risk of normalising extremist ideas. Automated recommendation systems, for instance, may inadvertently guide users toward progressively extreme virtual spaces without their conscious awareness.

The technical architecture of decentralised metaverses further complicates regulatory and oversight efforts (Effing, 2024; Goldberg and Schär, 2023). Unlike conventional online games, where the framework is established and controlled by the developers, decentralised metaverses afford users extensive creative freedom to design private spaces with advanced multimedia capabilities. This includes the integration of videos, sounds, and interactive objects, facilitating the creation of engaging and varied content. Consequently, extremist groups could organise virtual exhibitions, film screenings, or even themed environments dedicated to revisionist histories for instance. These private domains can be carefully controlled to maintain confidentiality and secrecy, limiting opportunities for external monitoring or intervention. This autonomy presents a fertile ground for the development and refinement of manipulative strategies by extremist actors.

3.3 The Radicalization Trade-off: Audience Size vs. Active Proselytism

On first analysis, one could think that decentralised metaverses could offer many attractions for extremist groups and their propaganda operations.

Extremist groups can indeed be attracted to decentralised metaverses for several reasons related to their technical architecture and operating principles.

The fragmentation inherent in decentralised governance systems presents significant challenges in the timely management of problematic content within metaverse environments. Unlike centralised platforms where moderators or administrators can swiftly remove harmful material, decentralised systems often rely on community-driven decision-making processes, such as voting or consensus mechanisms, which can be inherently slow and cumbersome. This procedural latency provides malicious actors with a critical window of opportunity to disseminate extremist or even criminal content before any remedial action can be taken, thereby increasing the potential reach and impact of such material.

This governance fragmentation also manifests in a heightened resistance to content censorship. The distributed nature of these platforms necessitates coordinated agreement among multiple network nodes or stakeholders to remove undesirable content or ban offending users. As a consequence, unilateral interventions are rare or practically impossible, and requests to eliminate problematic content are often unsuccessful. This systemic inertia not only complicates moderation efforts but may also foster environments where extremist narratives can proliferate with relative impunity. Compounding these difficulties is the increased degree of pseudonymity and confidentiality afforded by decentralised metaverses. Users frequently interact under cryptographic wallet addresses or other anonymous identifiers, without the need to disclose personally identifiable information. This stands in contrast to many centralised platforms, which increasingly implement identity verification protocols — often under regulatory pressure — to safeguard vulnerable populations, including minors. The preservation of user anonymity in decentralised settings, while protective of privacy, simultaneously poses significant challenges for accountability and the enforcement of community standards, particularly in contexts where harmful or extremist content may be involved.

By choosing to invest in decentralised metaverses, extremist groups would free themselves from any external constraint and could develop universes to their liking by reserving them for users who have shown their credentials and protecting them from external scrutiny under the pretext of property protection. Propaganda and recruitment operations could then be conducted within the framework of "digital gated communities."

However, in doing so, extremist groups would sacrifice their strategy of following the audience where it is found. One can reasonably think that decentralised metaverses will be marginal in terms of audience compared to centralised metaverses. Adherence to a decentralised metaverse presupposes significant technical prerequisites and is based on financial motivation involving a good deal of speculation. It is therefore very far from what users whose motivation is essentially ludic seek and who look for ease of access and use, which are the characteristics of centralised metaverses. Even if extremist groups prove capable of gathering the necessary resources to design and maintain private spaces in decentralised metaverses, these radicalised universes will only attract a very limited number of users compared to the games that major centralised metaverse companies will be capable of bringing together. The main strategy of extremist groups in the Metaverse should therefore logically extend, adapting it as needed to the new opportunities offered by technology, the "Follow the Audience" strategy that has been tested successfully in the universe of MMORPGs.

4 Conclusion

Despite the resounding failure of Horizon Worlds, the Metaverse is indeed the future of the internet as we know it. The maturation of technological building blocks, the cost reductions that will follow, and the development of applications responding to needs, particularly ludic ones, will lead to the multiplication of Metaverse instantiations in diverse forms, both centralised and decentralised.

As an informational space, the Metaverse will rapidly become a field of confrontation where multiple influence, propaganda, or manipulation operations will be deployed. The new possibilities offered by technology will facilitate these operations and make them more dangerous due to increased user engagement and extended possibilities for emotion regulation.

To anticipate the manipulation strategies that will be implemented in the Metaverse, it is possible to build upon the experience acquired in the field of MMORPGs which largely prefigure what centralised metaverses will be. Despite the control exercised by the companies that design them, centralised ludic metaverses will be parasitised by extremist groups pursuing a "Follow the Audience" strategy and endeavouring, by exploiting game mechanisms, to convey, under the guise of humour particularly, extremist messages likely to seduce young users before attracting them to external forums where they will be subjected to indoctrination processes. These extremist groups will take advantage of new opportunities to create content, adapt it to different contexts, languages, or cultures, and make it evolve according to the reactions obtained whilst avoiding moderation devices if they exist. This strategy will be particularly pernicious and, given the ideological proximity of certain of the the main entrepreneurs of the field with these radicalised discourses, it should be the subject of active policy conducted by public authorities (particularly judicial) and democracy defence associations.

The creation of radicalised universes in decentralised metaverses appears as a spectacular but ultimately secondary threat. Certainly, technological progress will authorise certain extremist groups to construct universes in their image and they will be able to develop them all the more effectively as the tools at their disposal facilitate this creative work (from a technical and financial point of view particularly). These universes can also be constructed as private spaces accessible only by invitation or authorisation, which will shelter them from external scrutiny and judicial prosecution. However, just as the experience of creating online games by extremist groups has shown, this type of universe is not sought by the vast majority of users who prioritise the ludic dimension and not the ideological quest.

Overall, the question of information manipulation and extremist propaganda in the Metaverse is situated more in a relationship of adaptive continuity than in a complete rupture with our digital environment.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Adinolf S, Turkay S (2011) Controlling your game controls: Interface and customization. In: Proceedings of the 7th international conference on Games+ Learning+ Society Conference, pp 13–22

Aguerri JC, Santisteban M, Miró-Llinares F (2023) The enemy hates best? toxicity in league of legends and its content moderation implications. European Journal on Criminal Policy and Research 29(3):437–456

Allchorn W, Orofino E (2025) Policing extremism on gaming adjacent platforms: Awful but lawful? Frontiers in Psychology 16:1537460

Andrea K (2025) Gaming industry report 2025: Market size and trends. Udonis Mobile Marketing Agency

Basdevant A, François C, Ronfard R (2022) Mission exploratoire sur les métavers. Rapport officiel édité par le ministère de la Culture et le ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique, https://www vie-publique fr/rapport/286878-mission-exploratoire-sur-les-metavers p 25

Benaben F, Congès A, Fertier A (2025) A prospective vision of the evolution of immersive technologies: Towards a definition of metaverse. Technovation 140:103154

Boudreau K (2022) Beyond deviance: Toxic gaming culture and the potential for positive change. Critical Studies in Media Communication 39(3):181–190

Champion A, Frank R (2021) Exploring the "radicalization pipeline" on youtube. In: Terrorism risk assessment instruments. IOS Press, p 359–374

Condis M (2019) Hateful games: Why white supremacist recruiters target gamers and how to stop them. In: Digital Ethics. Digital Ethics. Rhetoric and Responsibility in Online Aggression. Taylor & Francis

Condis M (2021) Playing at racism: White supremacist recruitment in online video game culture. In: White Supremacy and the American Media. Routledge, p 246–274

Coutinho RW, Boukerche A (2023) When smart metaverse meets affective computing: Opportunities and design guidelines. IEEE Communications Magazine 61(10):46–52

Darras É (2020) Du rassemblement de tuning au rond-point des gilets jaunes. Actes de la recherche en sciences sociales 235(5):12–31

Davey J (2024) Extremism on gaming (-adjacent) platforms. Gaming and extremism: The radicalization of digital playgrounds pp 95–109

De Rossi LM, Abbatemarco N, Diaferia L, et al (2024) A metaverse typology: Exploring the role of immersivity and centralization in driving value creation. In: Proceedings of the Hawaii International Conference of System Science (HICSS)

Dendi NR (2025) Real-time content moderation in gaming platforms: Technical frameworks for child protection. Journal of Computer Science and Technology Studies 7(9):01–08

Donner F (2024) Structures that tilt: Understanding "toxic" behaviors in online gaming. New Media & Society p 14614448241270446

Effing R (2024) Will the metaverse be out of control? addressing the ethical and governance implications of a developing virtual society. Digital Government: Research and Practice 5(3):1–15

Ehrsson HH (2020) Multisensory processes in body ownership. Multisensory perception pp 179–200

Ferguson CJ, Glasgow B (2021) Who are gamergate? a descriptive study of individuals involved in the gamergate controversy. Psychology of Popular Media 10(2):243

Ghosh A, Hassija V, Chamola V, et al (2024) A survey on decentralized metaverse using blockchain and web 3.0 technologies, applications, and more. IEEE Access

Goldberg M, Schär F (2023) Metaverse governance: An empirical analysis of voting within decentralized autonomous organizations. Journal of Business Research 160:113764

Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7(1):2053951719897945

Grayson N (2022) How twitch took down buffalo shooter's stream in under two minutes. The Washington Post pp NA–NA

Hern A (2022) Online hate under scrutiny after buffalo shooter streamed massacre on twitch. The Guardian pp NA–NA

Karaarslan E, Yazici Yilmaz S (2023) Metaverse and decentralization. In: Metaverse: Technologies, Opportunities and Threats. Springer, p 31–44

Kilmer E, Aslan Z, Kowert R (2024) From avoidance to action: A call for open dialogue on hate, harassment, and extremism in the gaming industry. ACM Games: Research and Practice 2(2):1–17

Kilteni K, Maselli A, Kording KP, et al (2015) Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. Frontiers in human neuroscience 9:141

Kim DY, Lee HK, Chung K (2023) Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship. Journal of Retailing and Consumer Services 73:103382

Koehler D, Fiebig V, Jugl I (2023) From gaming to hating: extreme-right ideological indoctrination and mobilization for violence of children on online gaming platforms. Political Psychology 44(2):419–434

Kowert R, Kilmer E, Newhouse A (2024) Taking it to the extreme: prevalence and nature of extremist sentiment in games. Frontiers in Psychology 15:1410620

Lakhani S, Wiedlitzka S (2023) "press f to pay respects": an empirical exploration of the mechanics of gamification in relation to the christchurch attack. Terrorism and political violence 35(7):1586–1603

Lam KY, Yang L, Alhilal A, et al (2022) Human-avatar interaction in metaverse: Framework for full-body interaction. In: Proceedings of the 4th ACM International Conference on Multimedia in Asia, pp 1–7

Lamphere-Englund G, White J, et al (2023) The online gaming ecosystem: Assessing digital socialisation, extremism risks and harms mitigation efforts

League AD (2024) Hate is no game. Tech. rep., https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023

Ledwich M, Zaitsev A (2019) Algorithmic extremism: Examining youtube's rabbit hole of radicalization. arXiv preprint arXiv:191211211

Lee HW, Chang K, Uhm JP, et al (2023) How avatar identification affects enjoyment in the metaverse: the roles of avatar customization and social engagement. Cyberpsychology, Behavior, and Social Networking 26(4):255–262

Mamié R, Horta Ribeiro M, West R (2021) Are anti-feminist communities gateways to the far right? evidence from reddit and youtube. In: Proceedings of the 13th ACM Web Science Conference 2021, pp 139–147

Mattheis AA, Kingdon A (2023) Moderating manipulation: Demystifying extremist tactics for gaming the (regulatory) system. Policy & Internet 15(4):478–497

Munn L (2019) Alt-right pipeline: Individual journeys to extremism online. First Monday

Munn L (2023) Toxic play: Examining the issue of hate within gaming. First Monday 28(9)

Mystakidis S (2022) Metaverse. Encyclopedia 2(1):486–497

Nissen R, Avramov K, Roberts J (2021) White rex, white nationalism, and combat sport: The production of a far-right cultural scene. Journal of Illiberalism Studies 1(2):19–37

Pervez F, Shoukat M, Usama M, et al (2024) Affective computing and the road to an emotionally intelligent metaverse. IEEE Open Journal of the Computer Society 5:195–214

Pira F (2024) The technological revolution, the metaverse and artificial intelligence. Addiction & Social Media Communication 1(1):3–12

Plans Vilaseca J (2024) Pixelated sexism: Gta v's dangerous influence on gender roles. Available at SSRN 5006787

Ribeiro MH, Ottoni R, West R, et al (2020) Auditing radicalization pathways on youtube. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 131–141

Robinson N, Whittaker J (2020) Playing for hate? extremism, terrorism, and videogames. Studies in Conflict & Terrorism pp 1–36

Sampedro PP, Grassi NB, Zitkus E, et al (2023) Experiences through customizations: looking beyond game's mechanics. Observatorio de la Economia Americana 21(11):19929–19950

Schlegel L (2021) Positive play: Can gamification support p/cve measures? JEX Journal Exit Deutschland Mars

Schlegel L, Kowert R (2024a) Gaming and extremism: The radicalization of digital playgrounds. Taylor & Francis

Schlegel L, Kowert R (2024b) Introduction: Extremism in digital gaming spaces. In: Gaming and Extremism. Routledge, p 1–8

Schlegel L, Wiegold L, Winkler C, et al (2025) Exploring the digital extremist ecosystem: a preliminary analysis of hateful posts on mod db. Frontiers in Psychology 15:1502098

Schöbel S, Karatas J, Tingelhoff F, et al (2023) Not everything is a metaverse?! a practitioners perspective on characterizing metaverse platforms. In: Proceedings of the Hawaii International Conference of System Science (HICSS)

Silva JPN, Valadares GC, Pedrosa G, et al (2023) Gender imbalance in mmorpg: the case of world of warcraft in brazil. Feminist Media Studies 23(1):289–305

Sorell T, Kelsall J (2025) Violent video games, recruitment and extremism. Criminal Justice Ethics 44(1):1–24

Steinkuehler C, Squire K (2024) Introduction to videogames and the extremist ecosystem. "An excellent analysis of gaming and extremism, which highlights some of the most complex challenges in tackling abuse of a constantly-changing technology space A perfect introduction for a novice and expert alike"–Charley Gleeson, Open-Source Intelligence Analyst at Tech Against Terrorism, UK p 9

Tassi P (2025) Tassi, paul. « meta's awful 'horizon worlds' ad helps explain 70 billion metaverse loss. Forbes 17 février

Vysotsky S (2022) Mainstreaming hate: Far-right cultural strategies

Wang J, Makowski S, Cieślik A, et al (2023) Fake news in virtual community, virtual society, and metaverse: A survey. IEEE Transactions on Computational Social Systems 11(4):4828–4842

Wells G, Romhanyi A, Reitman JG, et al (2024) Right-wing extremism in mainstream games: A review of the literature. Games and Culture 19(4):469–492

Williams JP, Kirschner D, Deterding S (2024) Sociology and role-playing games. In: The Routledge Handbook of Role-Playing Game Studies. Routledge, p 243–260

A Socioinformatic Analysis of Community-based Fact-Checking: Why Community Notes is Not as Effective as X Claims

Julian M. Pelloth[1], Lena Pölzer[1], Katharina A. Zweig[1]

[1] Department of Computer Science, University of Kaiserslautern-Landau.

Corresponding authors. E-mails: julian.pelloth@edu.rptu.de; dem69pod@rptu.de; zweig@cs.uni-kl.de

Abstract

The Community Notes developed by X aim to fight misinformation by enabling users to add context in the form of notes beneath posts. The system employs an algorithm using a modified matrix factorization to determine which notes are displayed. It publishes notes that receive agreement on helpfulness from raters who, according to their past rating behavior, have different "points of view". Our analysis of publicly available Community Notes data reveals that only a small fraction of notes ultimately receives a "helpful" or "not helpful" status, while the majority remain in the intermediate status of 'needs more ratings'. This shows that Community Notes is ineffective in adding notes and, therefore, lacks effectiveness in moderating posts and preventing the spread of misinformation.

To investigate the underlying reasons for this phenomenon, we apply a socioinformatic analysis. This method analyzes socioinformatic systems formed by interactions of social actors and their motivations with a software system. These interactions are modeled using a so-called effect network, allowing us to assess whether the ineffectiveness of the Community Notes emerges from how social actors engage with the system.

Our analysis explains the ineffectiveness of Community Notes in adding notes to (misleading) posts with the assumption that most raters are rating the note whether they align with their beliefs and not according to their helpfulness. The lack of notes undermines the goal of Community Notes to prevent the spread of misinformation. Finally, we identify several research questions that arise from our analysis.

Keywords

Countering Misinformation, Community Notes, Socioinformatic Analysis, Disconfirmation Bias.

1 Introduction

Even though social media platforms have repeatedly been criticized for their role in the dissemination of misleading information, individuals increasingly rely on these platforms as news sources. The Digital News Report of 2023 by the Reuters Institute shows that since 2018, there has been a notable shift in how people access online news. On average, 30% of people across countries prefer to get their online news through social media, surpassing the 22% who choose direct access to news websites or apps (Reuters Institute for the Study of Journalism, 2023).

Given the increasing reliance of individuals on social media as a primary source of online news, countering misinformation on these platforms becomes crucial. This issue becomes more urgent as more platforms rely on community-based fact-checking: X (formerly Twitter) introduced the *Community Notes* (formerly Birdwatch) in 2021.[9] In 2025, Meta announced their plans to replace their third-party fact-checking with a crowd-sourced Community Notes approach[10] and TikTok stated its plans to incorporate the so-called Footnotes[11], based on Community Notes, into its existing content moderation. This emphasizes the need for a detailed analysis of community-based fact-checking technologies.

Community Notes (CN) is a technology developed and deployed by X to combat misinformation. It enables users to add context, missing information, or sources in the form of *notes* beneath posts (Wojcik, Hilgard, Judd, Mocanu, Ragain, Hunzaker, Coleman, and Baxter, 2022; X, 2024). Contrary to traditional content moderation methods which typically involve the removal of posts by moderators, the CN prioritizes user participation and a "consensus-driven approach to truth"[12]. The notes can be rated by *contributors* as "helpful"[13], "somewhat helpful", or "not helpful". For a note to be displayed to all users, it has to be rated as "helpful" by raters with "different points of view". To determine the viewpoints of contributors and which note receives a "helpful" status, X employs an algorithm that utilizes a specialized matrix factorization and the past rating behavior.

We analyzed the publicly available CN data.[14] As seen in Figure 1, the analysis revealed that only approximately 16% of eligible notes ultimately receive a "helpful" (11%) or "not helpful" status (5%), while the majority remain in the intermediate status of "needs more ratings".[15] The findings indicate that CN is ineffective in adding notes and, therefore, lacks effectiveness in moderating posts and preventing the spread of misinformation.[16]

While X states in the documentation that they "set the thresholds to achieve a 'Helpful' status at 0.40, including less than 10% of the notes"[17], they do not claim that the system is designed to ensure that only ten percent of all notes are classified as "helpful". The decision of setting the threshold this way is not explained in the documentation. The current design of the algorithm (see section 3) could achieve more than 10% of notes with a "helpful" status. Despite this, we also aim to explain why there is such a high percentage of notes with a "needs more ratings" status, because this means that a lot of notes remain undecided and are neither classified as "helpful" nor "not helpful".[18]

We apply a socioinformatic analysis to investigate the underlying reasons for this phenomenon (Zweig, Krafft, Klingel, and Park, 2021). This method analyzes socioinformatic systems formed by interactions of social actors and their motivations with a software system. These interactions are modeled using a so-called *effect network*, allowing us to assess whether the ineffectiveness of the Community Notes emerges

[9] see https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation (accessed 27.05.2025).

[10] see https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/ (accessed 04.06.2025) and https://www.meta.com/technologies/community-notes/?srsltid=AfmBOopUQHPjbAyu5FvQqslS-ZMmR1W5HMLuvPcfaSCVkhsNKln6s76E (accessed 04.06.2025).

[11] see https://newsroom.tiktok.com/en-us/footnotes (accessed 27.06.2025).

[12] see https://www.youtube.com/watch?v=41a_QDhOOuk&t=9s&ab_channel=TheEconomicTimes (accessed 02.03.2024).

[13] To differentiate between the actual helpfulness of a note and the "helpful" ratings and status, the second is always put in quotes.

[14] We retrieved the Community Notes data on multiple occasions first on 22.08.2023 and last on 23.07.2025. The data is available under https://x.com/i/communitynotes/download-data (accessed 23.07.2025).

[15] To be eligible a note has to receive at least five ratings. All notes with less than five ratings were excluded from the data, as they only receive a "needs more ratings status". These are around 16% of all notes. When looking at all notes, the status distribution is similar: 8% "helpful", 4% "not helpful" and 88% "needs more ratings" (see appendix Figure 1). The proportion of "helpful" and "not helpful" note is only decreased by three and one percent, while the proportion of "need more ratings" is increased by four percent. Therefore, the impact of the ineligible notes are marginal.

[16] This analysis does not take into account that the public available data contained only a total number of 1.7 million notes by the 13th May 2025. A last data retrieval on the 23th July 2025 showed that the data contained 2 million notes, which is the total number of notes since the release of the Community Notes in January 2021. This poses another question if the Community Notes can really cover all misleading posts on X. However, this is not part of this analysis.

[17] see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#matrix-factorization (accessed 04.10.2025).

[18] Even if the 11% of "helpful" notes would be a result of the algorithm decision making process, it would be interesting why so many notes are labeled with "needs more ratings", which would not be the case if the other 89% of notes would all be labeled as "not helpful".

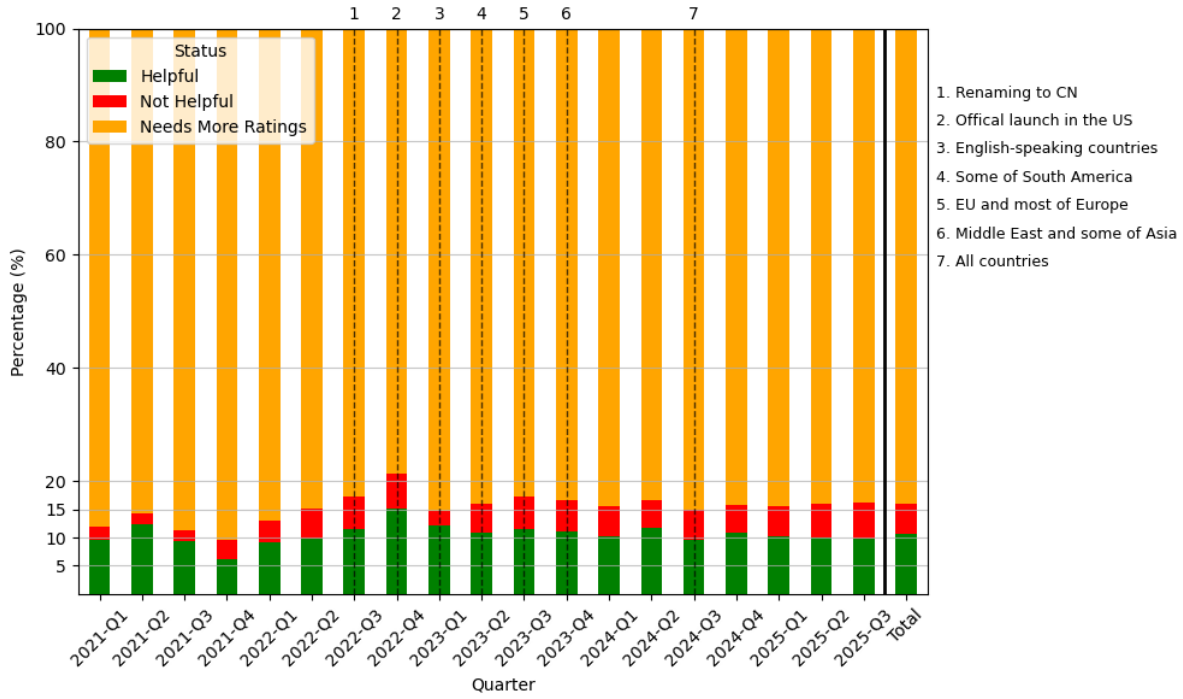


Fig. 1: The total distribution of the note status of the eligible notes shows that most eligible notes have the "needs more ratings" status (84%), around 11% receive the "helpful" status and 5% the receive the "not helpful" status. The status distribution stays approximately the same since the first quarter after the official launch. To be eligible for the note scoring algorithm a note has to receive at least five ratings, all notes with less than five ratings were excluded. For the distribution of the number of notes see appendix Figure 2 and for a distribution of the note status with all notes see see appendix Figure 1.

from how social actors engage with the system. For the purpose of the analysis, we choose the contributors that rate notes and X as the social actors. We assume that the contributors either aim to rate notes according to their actual helpfulness or their prior beliefs. X aims to reduce the number of misleading posts without notes. Overall, this method aims to analyze how the distribution of the note status is a result of the design choices for the CN algorithm and the incentives it provides for the contributors.

2 Socioinformatic Analysis

The socioinformatic analysis, according to Zweig et al. (2021), belongs to the field of socioinformatics, whose goal is to "analyze and - where possible - predict the effects of software and hardware systems on individuals, organizations, and society [Translation by the author (TbA)]"[19](Zweig et al., 2021, 46). Socioinformatics adopts the approach that the use of software changes human behavior, thereby creating technological consequences (Zweig et al., 2021).
According to the authors, the technical component and the social actors form a so-called *socio-technical system*. It consists of at least one human component and informatic system, and in which human behavior is shaped by the use of the software. Thus, it is called a *socioinformatic system*.[20] The exhibited changes in human behavior can lead to a *socioinformatic phenomenon* that are defined as an emergent phenomenon caused by the interaction of social actors with software that cannot be reduced solely to the behavior of humans or the informatic system (Zweig et al., 2021).

[19]"Das Ziel der Sozioinformatik ist es, die Auswirkungen von Software- und Hardwaresystemen auf Individuum, Organisation und Gesellschaft zu analysieren und – wo möglich – vorherzusagen"(Zweig et al., 2021, 46).
[20]The socioinformatic system is a subset of socio-technical systems that has a central software or hardware component (Zweig et al., 2021, 49).

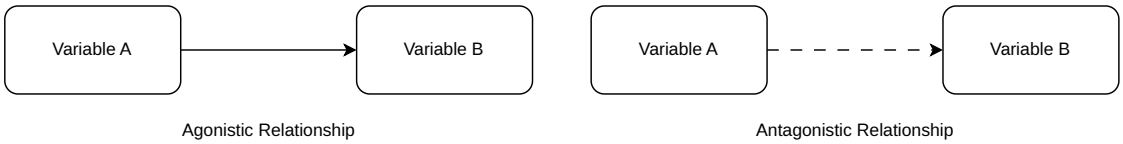


Fig. 2: The causal relationships in the effect network can be distinguished between agonistic and antagonistic relationships. The agonistic relationship is visualized with a solid line, while the antagonistic relationship is represented with a dotted line between two system variables (Zweig et al., 2021).

2.1 Phases of the Socioinformatic Analysis

The socioinformatic analysis consists of five phases, which will be explained in the following paragraphs, starting with the description of the technical foundation.

Description of the Technical Foundation

In the first phase, according to Zweig et al. (2021), the relevant informatics components that are either part of the phenomenon or respond to human action are identified and described (Zweig et al., 2021). For this description, it is crucial to understand the respective software or algorithms so thoroughly that it becomes clear how they influence the phenomenon and alter the incentive structures.[21]

Identification of the relevant Social Actors and their Motivations

In the second phase, the relevant social actors are first identified, and then their respective motivations are determined (Zweig et al., 2021). Social actors can be individuals, organizations, or even society. According to the authors, they generally belong to one of four groups: developers[22], operators[23], users of the software, and other social actors who interact with the informatic component, such as legislators.

Creation of the Effect Network

To explain how the software generates the socioinformatic phenomenon or which technological consequences could arise, Zweig et al. (2021) propose modeling the process as compactly as possible using an *effect network* (Zweig et al., 2021). An effect network visualizes the "causal dependencies between measurable properties or measurable behaviors of the system [TbA]"[24](Zweig et al., 2021, 99). Zweig et al. (2021) use the informal definition of causality by Pearl, Glymour, and Jewell: "A variable X is a cause of a variable Y if Y in any way relies on X for its value"((Pearl, Glymour, and Jewell, 2016), as cited in (Zweig et al., 2021, 14)). These measurable properties are called *system variables* and are represented as nodes. The causal relationships are represented as directed edges, illustrating how a change in the value of system variable A affects system variable B. Therefore, an effect network represents plausible causal relationships between system variables. They differentiate between two types of relationships: The *agonistic relationship* describes that an increase in the values of A increases the value of B and is visualized through a solid line, as seen in Figure 2. Contrary, the *antagonistic relationship* - a dotted line - describes that an increase in the value of A decreases the values of B (Zweig et al., 2021). To represent causal relationships, Zweig et al. (2021) require system variables to be quantifiable, meaning they must theoretically be assignable to a value, and an increase and decrease of the variables value must be plausible (Zweig et al., 2021). The motivations of social actors are visualized through so-called *motivational edges*. These are undirected edges connected to a system variable and indicate that a social actor seeks to influence the variable's value: A solid edge shows that the social actor wants to increase the value, while a dotted edge indicates a desire to decrease it.

The authors stipulate that a causal relationship between two variables should only be included in

[21] Zweig et al. (2021) distinguish between deliberately implemented incentives and those that unintentionally arise from developer decisions. Both types can influence implementation and result in intended as well as unintended technological effects. Specific incentives may be embedded in optimization functions that are not visible or transparent. The authors also note that it is challenging to identify unintended incentive structures before deploying a technological artifact (Zweig et al., 2021).

[22] Developers have direct influence over the code of the software (Zweig et al., 2021).

[23] Operators make the software available to others. They either purchased it or had it developed (Zweig et al., 2021).

[24] "Wie in Abschnitt 2.1.3 schon definiert, handelt es sich bei einem Wirkungsgefüge um ein sehr einfaches visuelles Modell von kausalen Abhängigkeiten zwischen messbaren Eigenschaften oder messbarem Verhalten des Systems"(Zweig et al., 2021, 99).

the effect network if it is plausible and, ideally, supported by a theory or empirical study (Zweig et al., 2021). For the analysis of the effect network, the existence, intensity, and plausibility of the causal relationships need to be verifiable.[25]

Analysis of the Effect Network
The system's dynamics are then examined, starting from the motivations of the social actors (Zweig et al., 2021). This involves investigating which system variable the social actor wants to increase or decrease according to their motivations. On the other hand, it must be analyzed which system variables the social actor can influence so that the desired changes occur through the causal relationships. Therefore, the resulting dynamics are analyzed by following the paths and loops that exist within the effect network and how they react to the motivations of the social actors (Zweig et al., 2021).

Further Research
In the last phase of the socioinformatic analysis, we aim to infer research questions that arise from the analysis of the effect network and should be further investigated.

In the next part, the first phase of the socioinformatic analysis will be executed, i.e. we will describe the technology behind the Community Notes.

3 Community Notes by X

In order to evaluate the Community Notes (CN) program from X, it is crucial to understand the technical foundations of the CN. The primary sources of information for the CN were the accompanying research paper (Wojcik et al., 2022), the documentation of the CN (X, 2024), and the available open-source code[26]. This chapter presents the first phase of the socioinformatic method, which is the description of the technical foundations.
The basic functionality of the CN is examined first.

3.1 Basic Functionality

Community Notes allow so-called *contributors* to add context to potentially misleading posts (X, 2024). Unlike traditional comments, they are displayed to users below the respective posts, ideally providing helpful context and credible sources, as depicted in Figure 3. The notes are not directly shown after they were written (X, 2024). They can just be seen by other contributors who can rate a note as "helpful", "not helpful", or "somewhat helpful". With their ratings contributors can assign each note tags that describe what attributes of the note are either helpful or not helpful.[27] These ratings are then utilized to compute the status of a note, which determines whether it is displayed beneath a post. The three possible statuses a note can obtain are "needs more ratings", "helpful", and "not helpful". Notes are initially categorized as "needs more ratings" until they accumulate enough ratings to surpass certain thresholds. In order to decide whether a note achieves the status "helpful" or not, the note ranking algorithm categorizes users based on their past rating behavior and displays only those notes that have been rated helpful by a sufficient amount of users with "*different points of view*"(X, 2024). The note ranking algorithm is described in the next section.

3.2 Note Ranking Algorithm

The upcoming section will examine the key details necessary for comprehending how the statuses of notes are computed. This analysis is of significant importance in understanding the precise categories that determine whether a note is deemed helpful or not. Such understanding is essential when evaluating the effectiveness of the approach in terms of filtering high-quality notes, particularly since the process relies solely on algorithms and is not monitored by humans. The amount of "helpful" ratings by contributors with so-called *different perspectives* is the main requirement for a note to be classified as helpful.

[25] According to Zweig et al. (2021), there exists no universally valid method to assess the quality of the effect network. However, they propose to check the motivations of the social actors, systems variables, and causal relationships according to their plausibility. Furthermore, they suggest using Occam's razor i.e. if a smaller effect network has the same explanatory power, it is better than a larger one (Zweig et al., 2021).
[26] see https://github.com/twitter/communitynotes (accessed 18.07.2025).
[27] The tags that can be assigned with the ratings are shown in Figure 4.

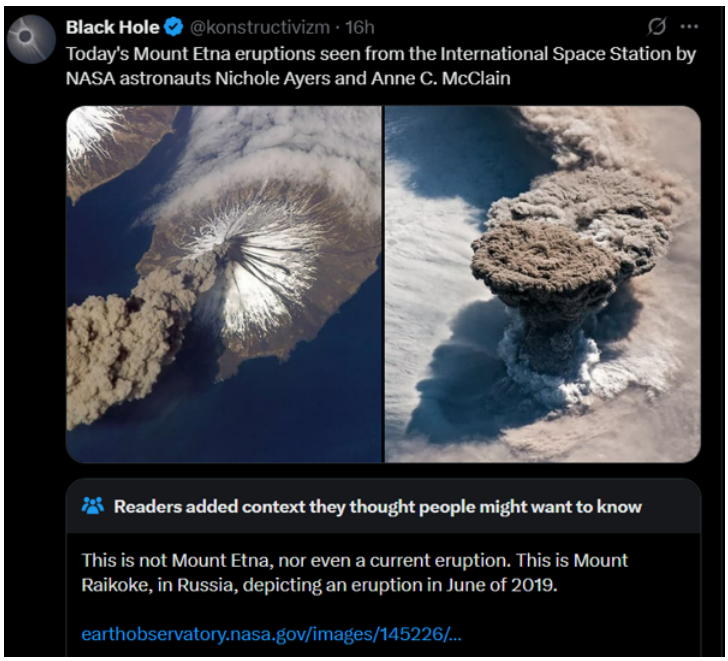

Black Hole @konstructivizm · 16h
Today's Mount Etna eruptions seen from the International Space Station by NASA astronauts Nichole Ayers and Anne C. McClain
Readers added context they thought people might want to know
This is not Mount Etna, nor even a current eruption. This is Mount Raikoke, in Russia, depicting an eruption in June of 2019.
earthobservatory.nasa.gov/images/145226/...

Fig. 3: Example of a note added to a post (see https://x.com/konstructivizm/status/1929665579901673658 (accessed 03.06.2025).

3.2.1 Matrix Factorization

The primary technique employed to determine whether a note is helpful or not is the matrix factorization, which, according to the documentation provided by X, is based on the work of Koren et al. (2009) (Koren, Bell, and Volinsky, 2009; Wojcik et al., 2022; X, 2024). The matrix factorization proposed by Koren et al. (2009) is used in recommender systems to provide users with fitting recommendations based on their item ratings. In this approach, a matrix encodes each user and each item, indicating whether a user has rated the item and how. Since this matrix is typically sparse, as most users only rate a small number of items, matrix factorization is employed to address this issue. The approach constructs a latent representation space for factors describing the items and the users. In this factor space, each item i is associated with a vector $q_i \in \mathbb{R}^f$, and each user u is associated with a vector $p_u \in \mathbb{R}^f$. These vectors aim to represent the characteristics of an item i and the preferences of a user u. Each rating is estimated by the multiplication of the respective item and user vector.[28] In order to learn these factor vectors, the model minimizes the regularized squared error via gradient descent on the set of known ratings (Koren et al., 2009).

The process of assigning statuses to notes follows a similar approach. As previously stated, the goal of the CN is to identify notes that are helpful to people with different points of view. According to X, this approach is based on the assumption that users with similar rating behavior have similar viewpoints.[29] Therefore, they conclude that "[i]f people who typically disagree in their ratings agree that a given note is helpful, it's probably a good indicator the note is helpful to people from different points of view"[30]. This is the reason why it is important for a note to not only be rated as helpful or unhelpful by many contributors, but they have to be from different perspectives to receive a "helpful" status.

The assessment of the perspective of the contributor is entirely based on the ratings of notes of the contributors. X employs a modified matrix factorization for the CN to determine the status of each note (Wojcik et al., 2022; X, 2024). The matrix factorization is trained with the given ratings[31] to find a representation space for the users (f_u) and notes (f_n) that aims to represent the affinity of certain users for certain features in the notes and what features a note has. X states that the "representation

[28] The estimate $\hat{r}_{ui} = q_i^{\mathrm{T}} p_u$ then indicates the overall interest a user u has in the characteristics of an item i because the dot product of two vectors yields a high value when both vectors align in the same direction, indicating a strong correlation between the user's preferences and the item's characteristics, corresponding with a high estimated rating (Koren et al., 2009).

[29] see https://communitynotes.x.com/guide/en/contributing/diversity-of-perspectives (accessed 03.07.2025).

[30] see https://communitynotes.x.com/guide/en/contributing/diversity-of-perspectives (accessed 03.07.2025).

[31] The ratings are mapped as follows: A "helpful" is a 1.0, "Somewhat helpful" 0.5 and "not helpful" rating maps to a 0.0 (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#helpful-rating-mapping (accessed 18.07.2025)).

space identifies whether notes may appeal to raters with specific viewpoints"[32]. A notable modification compared to Koren et al. (2009) is that the note and user representation space are one-dimensional[33], meaning that the categorization of the perspective of the user is based on a single value. Therefore, X claims that the representation space of the users and notes can capture the viewpoint of each user and note. In a two-party system, like in the USA, the user factor could map if a contributor leans more to one party or the other.
Since the algorithm aims to determine whether notes are broadly considered helpful by raters with different "viewpoints", the matrix factorization has to be extended as the proposed method by Koren et al. does not include a variable that aims to capture the helpfulness of a note. Therefore, the note ranking algorithm includes three additional parameters for the prediction of the rating (Wojcik et al., 2022; X, 2024). First, the global intercept term μ captures the overall tendency of all raters to rate notes higher or lower. Second, the user intercept term i_u represents a user-specific tendency in rating behavior whether contributors give out more or less "helpful" ratings to notes. Similarly, the third parameter, the note intercept term i_n, represents the note-specific bias and captures the intrinsic "helpfulness" of a specific note, independent of other factors. This parameter later determines whether a note is "helpful" or "not helpful". The predicted rating is expressed as:

$$\hat{r}_{un} = \mu + i_u + i_n + f_u \cdot f_n \tag{1}$$

The addition of the intercept terms aims to "capture the helpfulness of a note or rater that is not explained by viewpoint agreement"[34]. To learn these parameters, the model minimizes the regularized squared error via gradient descent on the set of known ratings.[35] A note receives a "helpful" status when it obtains a high note intercept score.
As seen in Equation 1, the note intercept can only be high when the multiplication of note and user factor receives a low value. This can only occur when the note is rated by contributors with lower and higher user factors. Figure 4 visualizes examples of the rating behavior of users with different factor vectors that lead to either a "helpful", "needs more ratings" or "not helpful" status. As seen in Figure 4a and Figure 4b, all users agree on the helpfulness of the note, which leads to the "(Not) Helpful" status. However, when only users with a lower user factor agree that the note is helpful while users with a higher user factor rate it as not helpful, the note obtains the "needs more ratings" status, depicted in Figure 4c. Figure 4 displays examples with distinct rating behavior. This does not necessarily happen in real life and a note can obtain a "helpful" status with a few "not helpful" ratings when there are still enough "'helpful" ratings by users with different perspectives. However, from the description of the algorithm it cannot be inferred when the contributors rating the note "helpful" are diverse enough or how many "helpful" ratings compared to "not helpful" ratings are needed to receive a "helpful" status. The matrix factorization is then used in different models within the note ranking algorithm. The different steps will be described next.

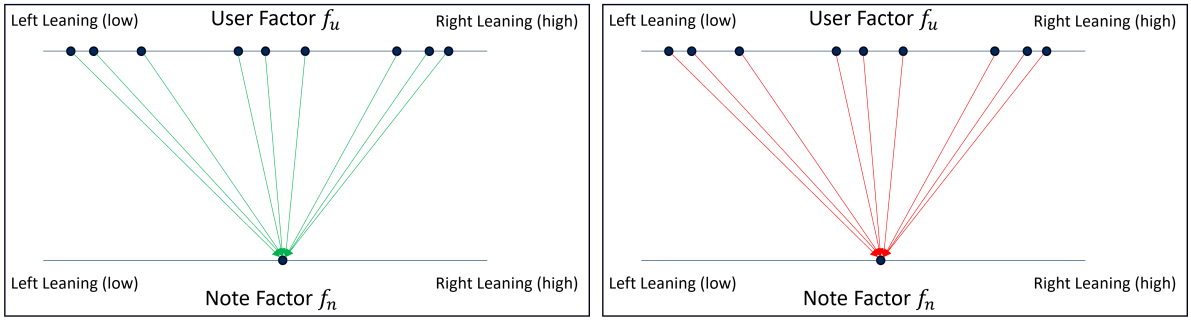
3.2.2 Algorithm Steps

According to the documentation provided by X, the note ranking algorithm[36], displayed in Figure 5, consists of two general steps, the prescoring and scoring:

Prescoring

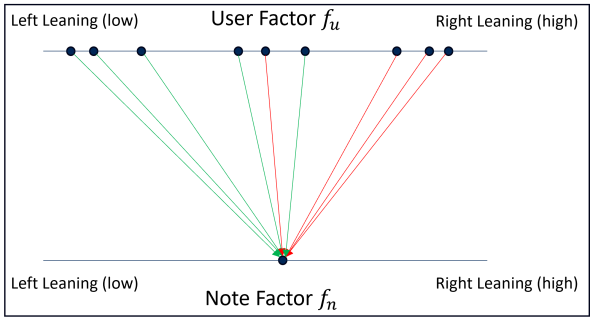
In the prescoring, the data is first pre-filtered to reduce some of the sparsity in the data. Therefore, only raters who have at least rated ten notes and notes who have at least five ratings are included in the data. Additionally, ratings from contributors with an anomalous rating behavior are merged.[37] Then, for multiple scorers that differ in what contributors are included in that data the model is trained on[38] and each runs the following algorithm steps separately. Additionally, there is the topic modeling, the main

[32] see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#matrix-factorization (accessed 18.07.2025).
[33] see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#matrix-factorization (accessed 18.07.2025).
[34] see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#matrix-factorization (accessed 18.07.2025).
[35] In general, the model is designed to minimize false positive predictions of note helpfulness and, therefore, prioritize precision over recall (for further reading see (Olson and Delen, 2008, 137-139)). In order to ensure that the model's predictions are highly reliable, even if it means sacrificing coverage of all actual helpful notes (Wojcik et al., 2022).
[36] For a detailed documentation of the note ranking algorithm see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes and for the open-source code see https://github.com/twitter/communitynotes/tree/main
[37] X calls this the *post selection similarity* (PSS) and the contributors with a high PSS are treated as a single contributor https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#agreement-safeguards (accessed: 18.07.2025).
[38] (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#multi-model-note-ranking (accessed 18.07.2025)). The *Core* model includes geographical areas where the CN are available longer, which is then expanded with the *Expansion* and



(a) Note with a "helpful" status

(b) Note with a "not helpful" status



(c) Note with a "needs more ratings" status

Fig. 4: The status of the note depends on the user factors of the contributors that rate the note. The green arrows denote a "helpful" rating, and the red arrows a "not helpful" rating. Figure (a) shows an example of the ratings for a "helpful" status, where contributors with different user factors agree that the note is "helpful". Similarly, as seen in (b), the note receives a "not helpful" status if contributors with different user factors agree that the note is "not helpful". However, if only contributors with a lower user factor agree that the note is "helpful", while contributors with a higher user factor rate it as "not helpful", it will obtain the "needs more ratings" status, depicted in (c).

matrix factorization (mf) model is fitted, and an intermediate status is assigned to each note according to their scores. Secondly, the *author and rater helpfulness scores* [39] are calculated, and raters with a low helpfulness score are removed from the data set. Then, a second mf model, called *Harassment-Abuse Tag-Consensus* model[40], is trained to determine whether a note is seen as harassment or abuse by raters with "different points of view", and the Author and Rater Helpfulness Scores are updated.
Additionally, topic modelling is used to for notes on specific topics.[41] Each post is assigned one of the topics[42] depending of a predefined list of seed words[43] In a first phase a multi-class logistic regression model is used to predict the post topics to possibly assign unassigned post to a topic. Then the topic model calculates the note ratings in the same way as the core model, with the exception that it does not apply the prescoring like the other model, as depicted in Figure 5.

ExpansionPlus model. The *Group* models focuses on non-english notes and assigns contributors based on their region, country, and language. The *CoreWithTopics* model contains within the core model which are assigned to a topic.

[39] For a detailed description of the Author and Rater Helpfulness Scores see https://communitynotes.x.com/guide/en/under-the-hood/contributor-scores.

[40] The Harassment-Abuse Tag-Consensus model is trained on the harassment and abuse tag, which raters can assign with their rating, and the "helpful" ratings. Instead of the "not helpful" and "helpful" ratings like it is done in the main matrix factorization model. Each rater how rated a note that is determined to be abusive as "helpful" receives a penalty on their rater helpfulness score (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#tag-consensus-harassment-abuse-note-score (accessed 18.07.2025)).

[41] see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#topic-modeling (accessed 18.07.2025).

[42] Currently, these are the "Unassgined", "UkaineConflict", "GazaConflict", "MessiRonaldo", and "Scams" (see https://github.com/twitter/communitynotes/blob/main/scoring/src/scoring/enums.py (accessed 18.07.2025)).

[43] For the lists of the seed words see https://github.com/twitter/communitynotes/blob/main/scoring/src/scoring/topic_model.py (accessed: 18.07.2025)).

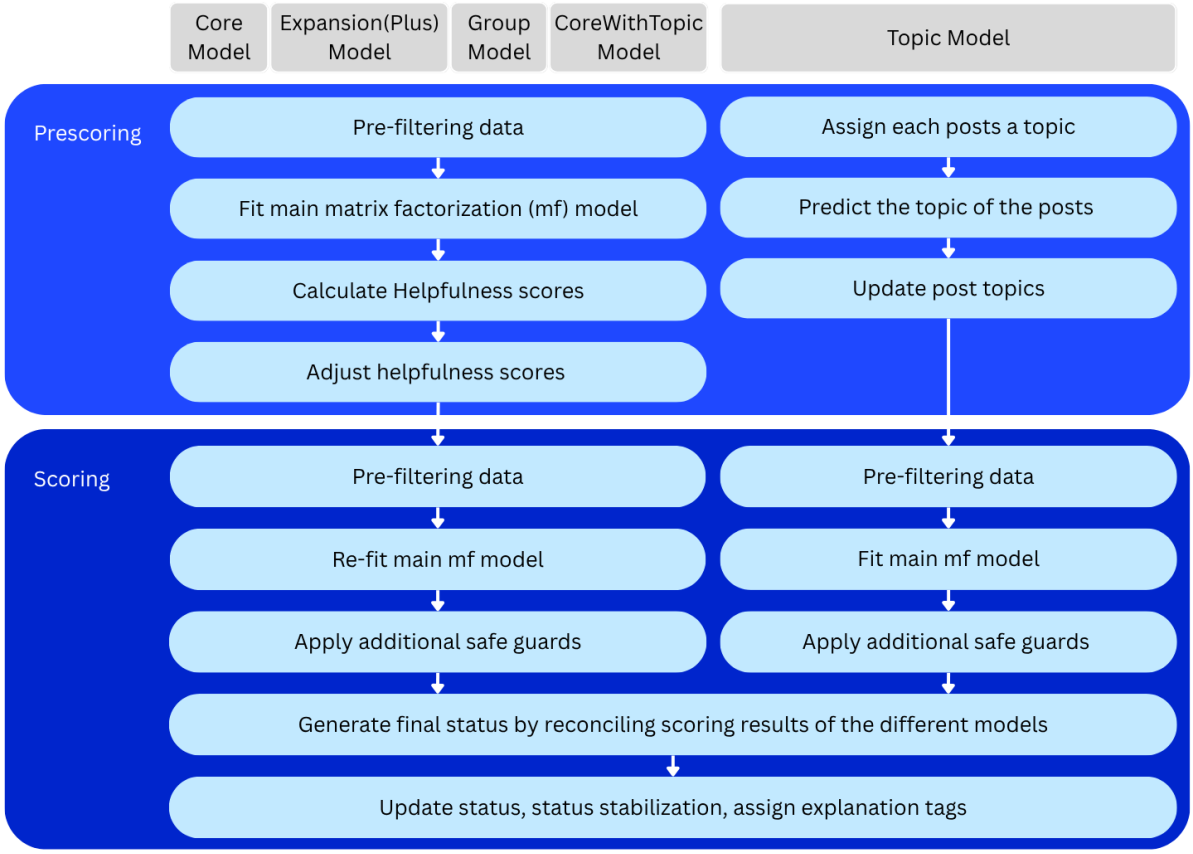


Fig. 5: The essential steps of the note ranking algorithm of the Community Notes.

Scoring

The scoring phase then consists of five steps. At first, the data from the prescoring is used, and for the newest notes and ratings data, the pre-filtering from the prescoring is applied. The main mf model is re-fitted with the filtered rating data set. Then additional safeguards are applied.[44]. The *note diligence* mf model, a third mf model, is trained with the goal to identify notes that are seen as "helpful" but have the issues with the their sources.[45] The main mf model is re-fitted with the addition of *pseudo raters* to calculate the lower and upper confidence bounds of all note intercepts.[46] Thirdly, the final status is obtained by reconciling the results of the different scorers.[47] Fourthly, the status labels of all notes that are less than two weeks old are updated, and for notes older than two weeks, the status is

[44]There are additional rules that are applied like the baseline intercepts for rater engagement and independence and additional steps in the algorithm (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#agreement-safeguards; https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#tag-outlier-filtering, https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#additional-filtering-for-incorrect-tags, https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#additional-not-helpful-scoring (accessed 23.07.2025)).

[45]The note diligence model applies the matrix factorization on the rating tags "Sources not included or unreliable", "Sources do not support note" and "Incorrect information". Furthermore, it learns a weight that categorizes whether rates are able to identify "helpful" notes that have reliable sources and increases the impact of the raters with a greater weight (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#note-diligence-score (accessed: 18.07.2025)).

[46]In this step "helpful" and Not Helpful" ratings of pseudo-raters that have the maximum or minimum user factor, or a factor of zero are added. In addition to the status thresholds a note is determined as "not helpful" when it a upper confidence bound of -0.04 (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#agreement-safeguards (accessed: 18.07.2025)).

[47]While all models calculate a status for the notes there are specific criteria which model is used to determine the final status. At first the core model is used when the note is also ranked by the Expansion or ExpasionPlus model. The group model can only promote a note from the "needs more ratings" to "helpful" status, under the condition that is receives a note intercept from the Expansion or Core model is between 0.3 and 0.4 (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#multi-model-note-ranking (accessed: 18.07.2025)). The topic models can only prevent notes from receiving a "helpful" status. This is the case when the note is not determined as "helpful" or when the note is "too aligned with a single perspective"(see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#topic-modeling (accessed 18.07.2025)).

stabilized.[48] In the last step, the notes get two explanation tags assigned according to their status.[49]

In the next phase, the involved social actors and their motivations will be identified.

4 Social Actors and their Motivations

In the second step of the method, the relevant social actors and their motivations are identified and outlined to understand the underlying dynamics of the socio-technical system comprised around the Community Notes. The main idea is that motivations of social actors are driving system variables who's change will then alter other system variables and thereby create the dynamics.

4.1 Company X

The first social actor is the company X, which developed the CN and deployed it on the social networking platform X. It has the same name as the company. According to the documentation provided by X (X, 2024), their primary motivation is to reduce the spread of posts containing misleading content. However, X only wants to get directly involved if a post violates the X guidelines. Otherwise, users should give additional context, which, according to X, is in the interest of freedom of speech.[50] Therefore, X is not involved in the note-writing and rating process. X intervenes when a note violates the guidelines.[51] The goal of X is to only classify notes as "helpful" if raters with different viewpoints agree on a note's helpfulness. This is the key paradigm of the algorithm, reflecting Elon Musk's aim for a "consensus-driven approach to truth"[52], as it has been highlighted in the analysis of the technical foundations.

4.2 Raters

The second social actors are the raters, who are contributors that rate notes. We can identify different motivations for the raters to rate notes. For this analysis, we categorize them into two different groups: *Fact-based Raters*, *Biased Raters*, and *Manipulative Raters*:

- **Fact-based Raters** are contributors who aim to rate notes according to their actual helpfulness because they want to prevent misleading posts from spreading without a note. They could, for example, assess the helpfulness of the note according to Fact-Checking Principles, like it is proposed by the International Fact-Checking Network.[53]
- **Biased Raters** are contributors who aim to rate notes according to their actual helpfulness but subconsciously reinforce their own point of view. This means that a contributor would rate a note according to their viewpoint because of the disconfirmation bias. According to Edwards and Smith (Edwards and Smith, 1996), disconfirmation bias describes the phenomenon that a person judges an incompatible argument as weak, while a compatible argument is more inclined to be accepted or judged as strong. Therefore, the notes that oppose their viewpoint are rated as "not helpful", and those that align are rated as "helpful". The disconfirmation bias is congruent with how humans assess information (Lewandowsky, Ecker, Seifert, Schwarz, and Cook, 2012; Schwarz, Newman, and Leach, 2016).
- **Manipulative Raters** want to actively reinforce their point of view or are actually hired to push specific agendas. Therefore, the contributor acts with a conscious decision to rate everything in favor of (or against) a specific viewpoint as "(Not) Helpful".

4.3 Environment

The remaining social actors are seen as part of the environment. This includes contributors who write notes (writers). This analysis focuses on the impact of the raters on the note status. While writers can

[48] The note status is finalized when it is two weeks old and the note status was determined by the Core, Expansion or Group models. Otherwise the note status can still change. According to X this is done with the intend to keep helpful notes from loosing their "helpful" status over time. Therefore, they are not affected by fluctuations of their ranking when the older data becomes a small fraction of the data set and not be affected by changes in the note ranking algorithm (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#status-stabilization (accessed 18.07.2025)).

[49] The explanation tags should explain why the note received a "helpful" status. The note receives the two explanation tags were selected the most by the contributors (see https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#determining-note-status-explanation-tags (accessed: 18.07.2025)).

[50] see https://help.x.com/de/resources/addressing-misleading-info (accessed 27.05.2025).

[51] see https://help.x.com/en/using-x/community-notes (accessed 27.06.2025).

[52] see https://www.youtube.com/watch?v=41a_QDhOOuk&t=9s&ab_channel=TheEconomicTimes (accessed 02.03.2024).

[53] see https://ifcncodeofprinciples.poynter.org/the-commitments (accessed 30.12.2024).

have an impact when they write a correct or misleading note or when it tends to be either right- or left-leaning, they do not influence the motivations underlying the rating behavior of the notes. The same applies to X-users who either write or react to notes, whose behavior does not influence the general rating behavior of the raters. Similarly, lawmakers and advertisers are seen as part of the environment because they do not have an immediate impact on the Community Notes and the rating behavior of the contributors.
In the next part, the effect network for the Community Notes process is constructed.

5 The Effect Network

5.1 Creation of the Effect Network

In the third phase, an effect network is constructed to model the exhibited phenomenon: The majority of notes remain in the "needs more ratings" status, while only around 11% of the eligible notes reach a "helpful" status. In order to better understand the reasons for this phenomenon, we constructed an effect network that explores the mechanisms arising from the motivations of raters and the motivation of X, coupled by causal relationships between system variables. The first step is thinking about the most import system variables that describe the socio-technical systems status.

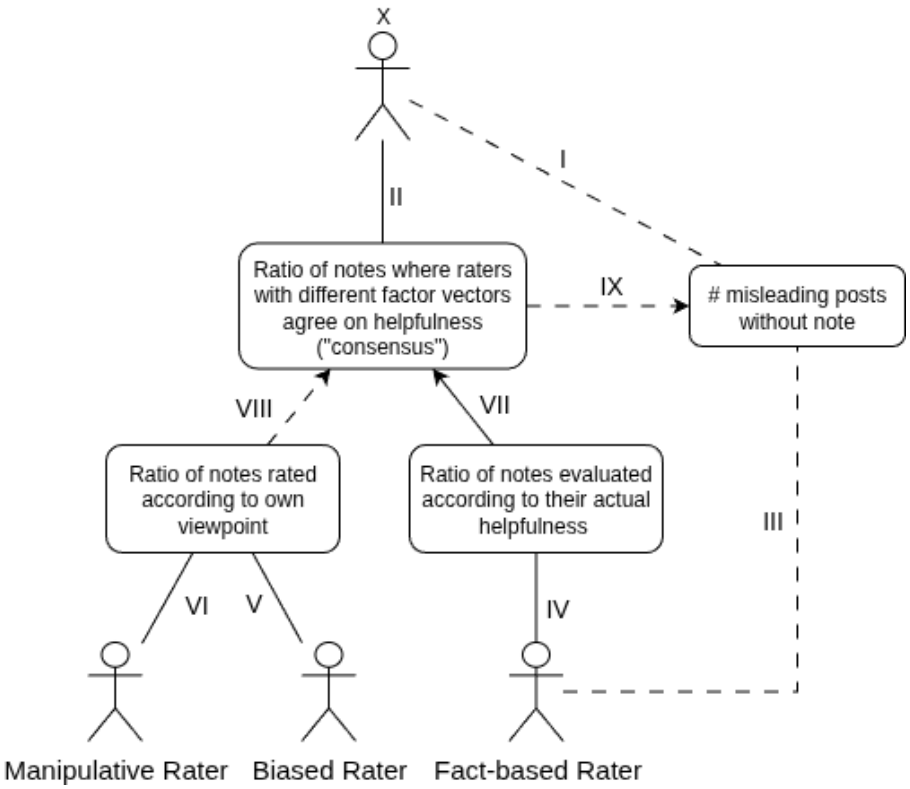


Fig. 6: The effect network shows the process of how the rating behavior impacts the number of notes that are shown on misleading posts.

Here, we only provide a short description of the main parts of the effect network while a full description of all system variables and edges is provided in the appendix.[54] In this effect network, the edges I to VI represent the motivational edges of the social actors that were described in the previous section. The edges VII to IX describe the impact of the rating behavior on the overall process: An increase in the *ratio of notes evaluated according to their actual helpfulness* results in an increase of the *ratio of notes where raters with different factor vectors agree on the helpfulness* (Edge VII) because then a high number of people agree that the note is either helpful or unhelpful. The note will finally receive the respective status. Contrary, an increase in the *ratio of the notes that are rated according to the viewpoint of the*

[54] For the full description of the system variables see Figure 6 and for the edges see Figure 6.

rater results in a decrease of the *ratio of notes where raters with different factor vectors agree on the helpfulness* (Edge VIII). This assumption is supported by the fact that contemporary societies are highly polarized (Barber, McCarty, Mansbridge, and Martin, 2015; Carothers and O'Donohue, 2019; Herold, Joachim, Otteni, and Vorländer, 2024; Pew Research Center, 2014). As a result, when the ratings are primarily based on their alignment with the raters' perspectives, there are fewer notes where raters with different factor vectors agree on their helpfulness.
Edge IX displays that an increase in the number of notes where raters agree on the helpfulness leads to a decrease in the *number of misleading posts without a note*, as more notes will be displayed on X. It should be noted that the assumption here is that the displayed notes are indeed helpful.
The following part will analyze the constructed effect network.

5.2 Analysis of the Effect Network

The effect network shows that for the Community Notes (CN) to work, a sufficient number of contributors need to rate the notes according to their actual helpfulness. This will lead to the consensus required for a note to receive a "helpful" or "not helpful" status. As Wojcik et al. (2022) state, they assume that "contributors rate notes as 'Helpful', 'Not helpful', or 'Somewhat helpful', based on whether the note provides helpful context or clarification about the Tweet or the issues it raises"(Wojcik et al., 2022, 1). This assumption is grounded in their belief that "regular people can valuably contribute to identifying and adding helpful context to potentially misleading information"[55] and is inscribed in the values that should guide the behavior of the contributors.[56] X is not naive in believing that contributors rate notes according to their actual helpfulness. As X states in the challenges of the CN, they are aware of the possibility of manipulation and biased outcomes.[57] Furthermore, the note ranking algorithm relies on previous disagreements between contributors in their ratings.[58]
X's statements imply an expectation that the majority of contributors are Fact-based Raters and not Biased or Manipulative Raters. However, if this were the case, we would expect more notes to reach a consensus on their helpfulness, which is not the case as only 11% of eligible notes receive a "helpful" status, whilst around 84% remain in the "needs more ratings" status.
The data appears to suggest that a majority of the contributors are Biased or Manipulative Raters rather than Fact-based Raters. Therefore, there are more ratings dependent on their prior beliefs. This phenomenon can in part be attributed to the disconfirmation bias, meaning that even though contributors may aim to rate notes according to their actual helpfulness, they are influenced by the bias and subconsciously follow their beliefs when rating notes. The biased behavior of the contributors can be observed in the open-source data of the Community Notes, as Figure 7 depicts.[59] It shows that as the distance between the Rater Factor and the Note Factor, which measures how closely the rater's viewpoint and that of the note align, increases the number of "helpful" ratings decreases. Specifically, "Not-Helpful" ratings tend to occur when there is a greater mismatch between the rater's and the note's viewpoint, as evidenced by the higher mean distance and a greater density at larger distances. Conversely, "helpful" ratings are associated with a much smaller distance, where viewpoints are closely aligned, leading to a higher concentration of low-distance occurrences. Under the assumption that the note and user factor approximately capture the viewpoint of the raters, it can be deduced that contributors primarily rate notes as helpful if their viewpoint aligns with that of the note. This is supported by Allen et al. (2022), who indicate that contributors are more likely to rate co-partisan[60] notes as "helpful" while having a lower likelihood of rating a counter-partisan note as "helpful" (Allen et al., 2022).
In summary, the effect network provides an explanation for the described phenomenon, where the majority are assigned a "needs more ratings" status, and only a small proportion of notes receive a "helpful" status. The following part will look into further research questions that arise from the analysis.

[55] see https://communitynotes.x.com/guide/en/about/introduction (accessed: 01.07.2025).
[56] see https://communitynotes.x.com/guide/en/contributing/values (accessed: 01.07.2025).
[57] see https://communitynotes.x.com/guide/en/about/challenges (accessed: 01.07.2025).
[58] see https://communitynotes.x.com/guide/en/contributing/diversity-of-perspectives (accessed: 01.07.2025).
[59] To obtain values such as rater and factor scores, the data was processed using the CN Algorithm, which is publicly available on GitHub. The figure was created using data that was downloaded on 19 May 2024.
[60] Allen et al. (2022) categorized contributors as either Democrat or Republican through the accounts the user followed (Allen, Martel, and Rand, 2022).

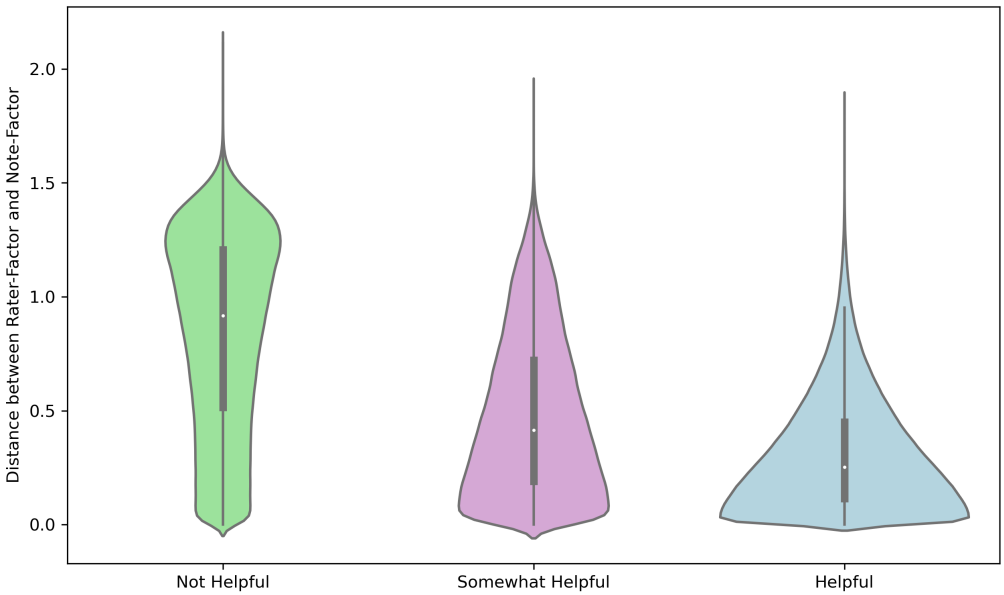


Fig. 7: The figure shows the distribution of the distance between the Note Factor and User Factor value, representing how much the *viewpoint* of the note and user align for each option on how a note can be rated. It includes all users and raters that are included in the scoring process, and receive a user and note factor. The distance is the calculated by subtracting the absolute values of the user and note factors.

5.3 Limitations and Further Research

The analysis shows that the high (low) number of notes with a "needs more ratings" ("helpful") status can be explained by the biased behavior of the contributors and by the design of the CN. This gives a first indication that the CN is prone to behavioral bias, which could hinder the publication of helpful notes. However, this is solely an indication that raises several points that need to be further studied.
The design of the algorithm and the rating behavior could result in the possible consequence that only uncontroversial notes actually receive a "helpful" status, while all notes on a controversial topic remain unseen by all users. We did not analyze the content of the notes. This could reveal what types of notes are published: Do notes on controversial posts to receive a "helpful" status or are most published notes uncontroversial posts, which our analysis would suggest.
Allen et al. (2022) further found a similar behavior for note writing, where counter-partisan posts are judged as more misleading than co-partisans (Allen et al., 2022). The concept of the disconfirmation bias would apply to the note writers, which could imply that the written notes are biased in the selection of the post, for which it was written, and the content of the note. However, this would have to be analyzed in more detail, e.g., by a separate effect network.
A third issue is the possible manipulation of the Community Notes. According to the documentation, X primarily focuses on preventing the manipulation of misleading notes that receive a "helpful" status.[61] However, we would add that the CN can be manipulated with many "not helpful" ratings to create an artificial dissent in the ratings of the note to prevent it from receiving a "helpful" status. This would also undermine the effectiveness of the CN and could be much harder to detect as most notes remain with a "needs more ratings" status.
Furthermore, we have identified discrepancies between the terminology used by X and the actual implementation of the CN algorithm. Especially how the "diversity of perspectives" is operationalized and how

[61] see https://communitynotes.x.com/guide/en/about/challenges (accessed 02.07.2025) and https://communitynotes.x.com/guide/en/under-the-hood/guardrails (accessed 02.07.2025).

the "consensus" among raters is determined by the algorithm, raising the question of whether past rating behavior can capture differing viewpoints and if a meaningful consensus among diverse perspectives is being reached.

6 Conclusion

This work aimed to analyze the rating process of the Community Notes (CN) with the socioinformatic analysis. We demonstrated that the low number of notes that are displayed on X can be attributed to the biased rating behavior of the contributors. As a result, our analysis yielded several research questions that merit further investigation.

In this work, we did not address whether the number of written notes could cover the potential amount of misinformation that is written on X[62], nor if the CN are shown fast enough to be able to have an impact. Additionally, we did not examine the overall effectiveness of the CN as a tool to combat misinformation. Wojcik et al. (2022) claim that the CN can reduce the spread of misinformation (Wojcik et al., 2022). A review by Martel and Rand (2023) suggests that crowd-sourced fact-checking labels can reduce the belief and spread of misinformation. However, they perform worse than labels issued by fact-checkers. It remains uncertain whether laypeople are effective in correctly identifying misinformation (Allen, Arechar, Pennycook, and Rand, 2021; Godel, Sanderson, Aslett, Nagler, Bonneau, Persily, and A. Tucker, 2021; Martel, Allen, Pennycook, and Rand, 2024). Nevertheless, Drolsbach et al. (2024) found that CN increases the trust in fact-checking (Drolsbach, Solovev, and Pröllochs, 2024).

This analysis highlights why it is important to investigate technologies like the Community Notes: On the one hand, the note ranking algorithm is designed with certain assumptions and design choices, which we revealed during our analysis of the provided documentation and code. On the other hand, the reason why the CN is ineffective is not necessarily the technology itself. However, it is embedded within a social process.The way the CN are integrated into the social process and how contributors interact with them is the real reason why the CN are ineffective. Socioinformatics specializes in analyzing technology in the context of the social process and approaches to human behavior and can identify the assumptions that need to be tested.

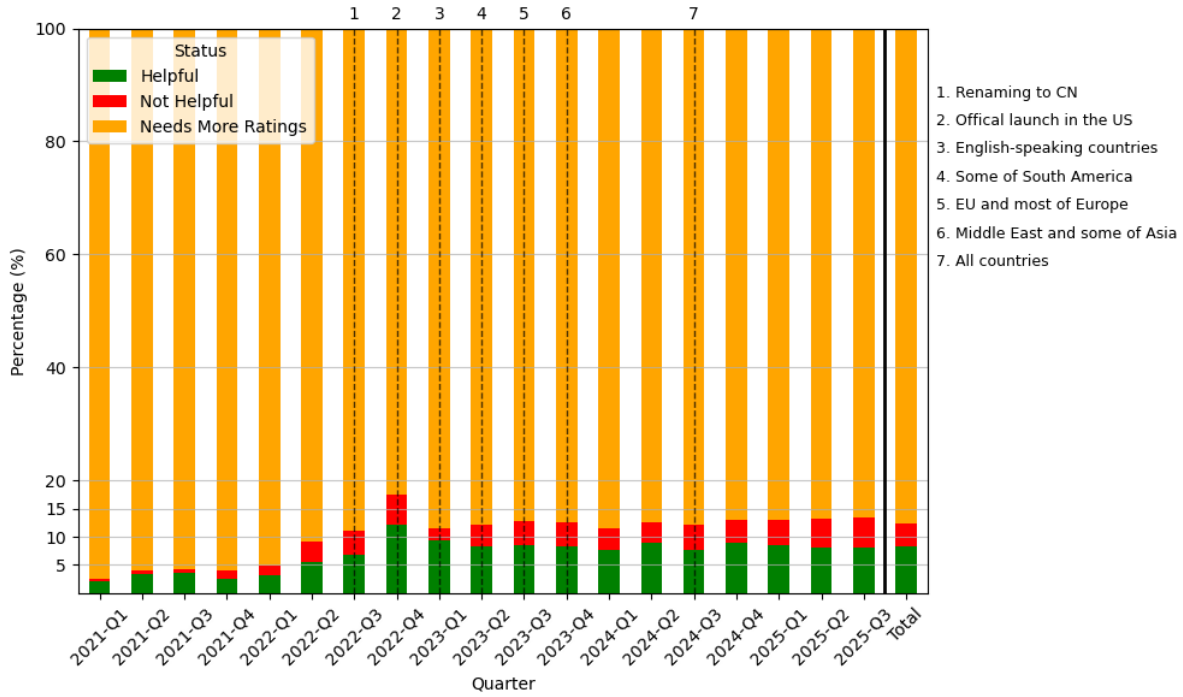
Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

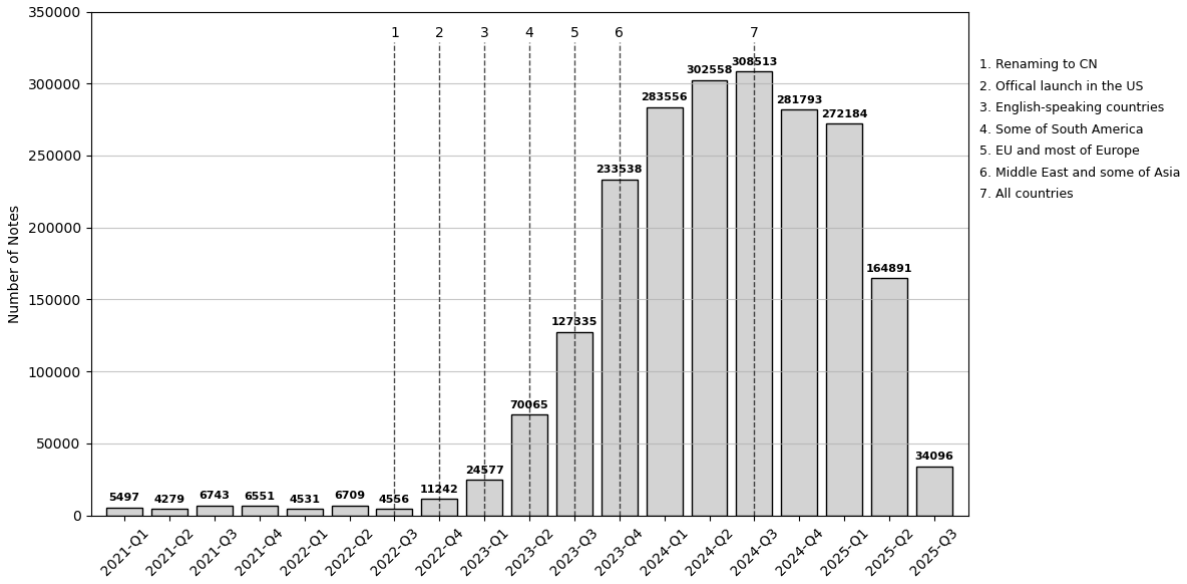
[62]The publicly available Community Notes data did only contain around 2 million notes, when retrieved on the 23.07.2025 and of those 16% were not scored, since they has less than five ratings.

Appendix

Note Status Distribution

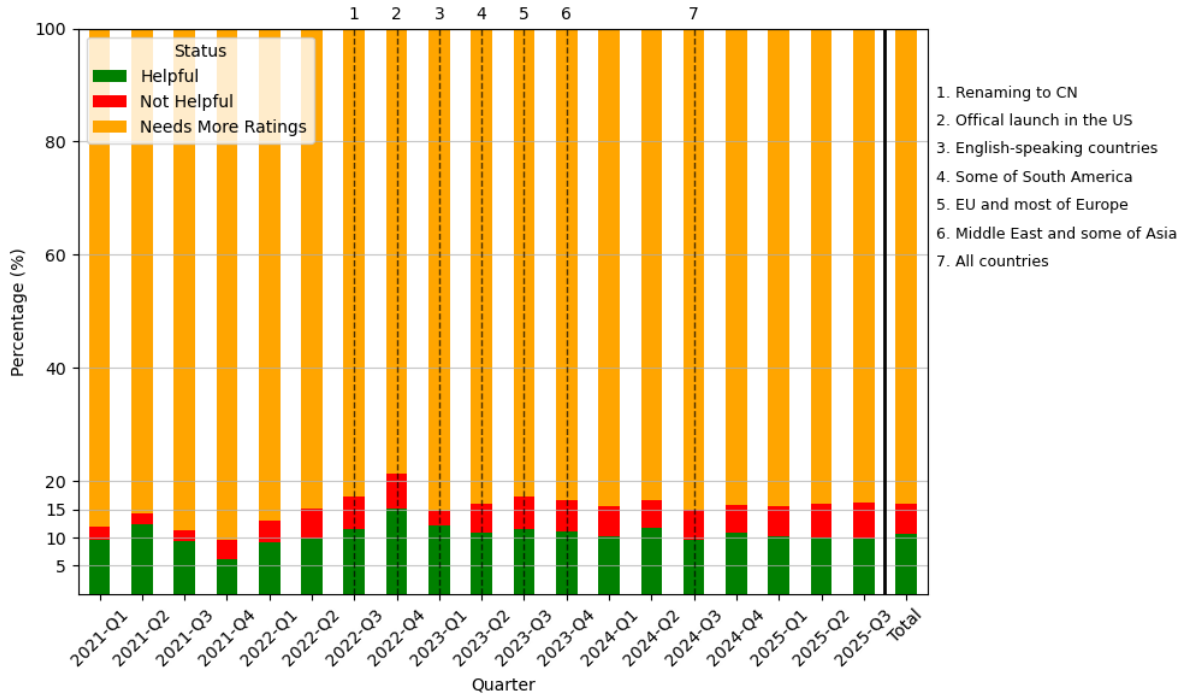


(a) Status distribution of all notes for each quarter the note was written in.

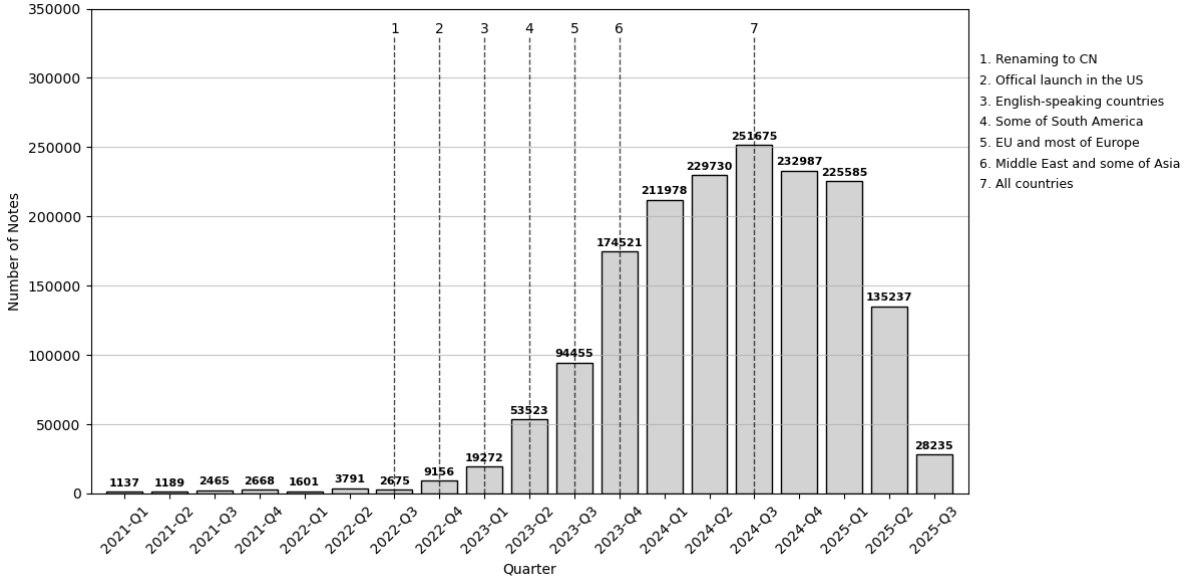


(b) Distribution of the number of notes that were written in each quarter. The data was collected during the third quarter of 2025.

Fig. 1: Note status distribution for all notes according to the quarter they were written in, with the addition of the overall distribution. The note status distribution stays approximately the same after Q1 2023, the quarter after the Community Notes were officially launched. The overall status distribution is 8% "helpful", 4% "not helpful" and 88% "needs more ratings" status. The number of notes that were written in each quarter can be seen in (b) and shows a steady rise in the number of notes, with the expansion of the Community Notes. After it was available in all countries, the number of notes decreased slightly and with Q2 2025 decreased by 40%. The data was collected during Q3 2025. Therefore, this bar shows only the notes written until the data was collected. Separate to each plot is the legend for the timeline where the Community Notes were launched.



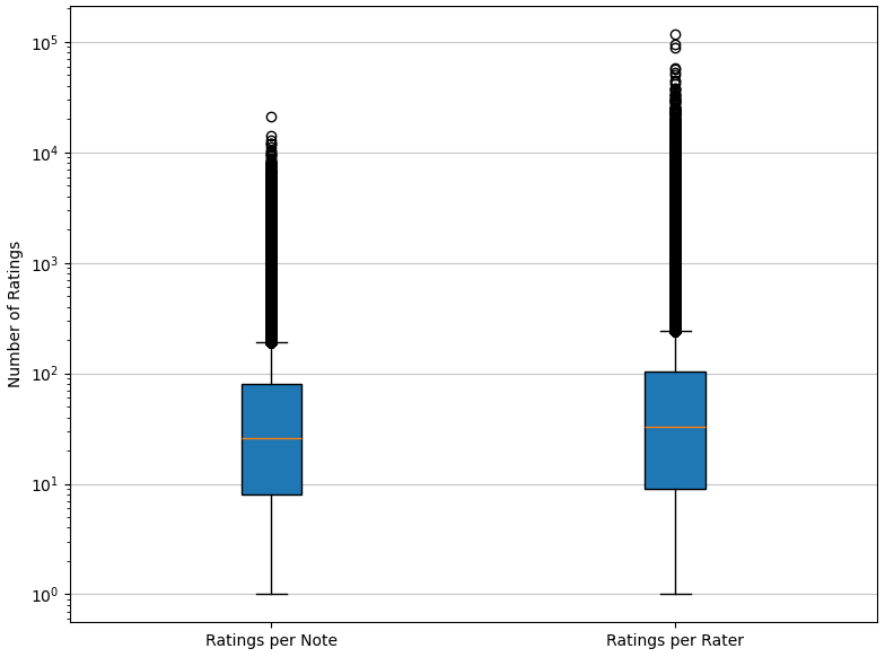
(a) Status distribution of the notes with five or more ratings for each quarter the note was written in.



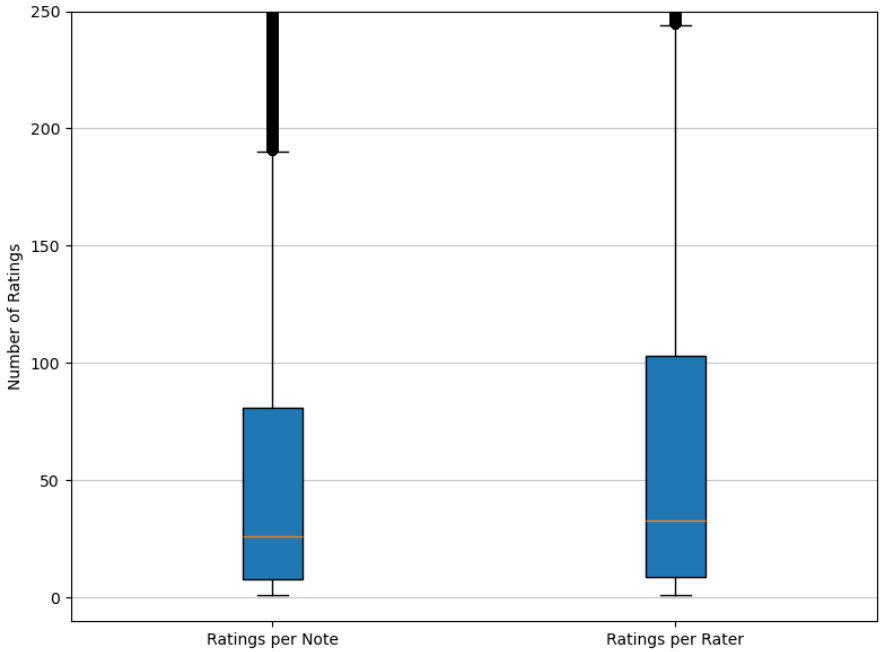
(b) Distribution of the number of notes with five or more ratings that were written in each quarter. The data was collected during the third quarter of 2025.

Fig. 2: Note status distribution for all eligible notes according to the quarter they were written in, with the addition of the overall distribution. All notes that have less than five ratings were removed. The note status distribution stays approximately the same after Q1 2023, the quarter after the Community Notes were officially launched. The overall status distribution is 11% "helpful", 5% "not helpful" and 84% "needs more ratings" status. The number of eligible notes that were written in each quarter can be seen in (b) and shows a steady rise in the number of notes, with the expansion of the Community Notes. After it was available in all countries, the number of notes decreased slightly and with Q2 2025 decreased by around 40%. The data was collected during Q3 2025. Therefore, this bar shows only the notes written until the data was collected. Separate to each plot is the legend for the timeline where the Community Notes were launched.

Ratings Analysis



(a) Ratings per users or notes.

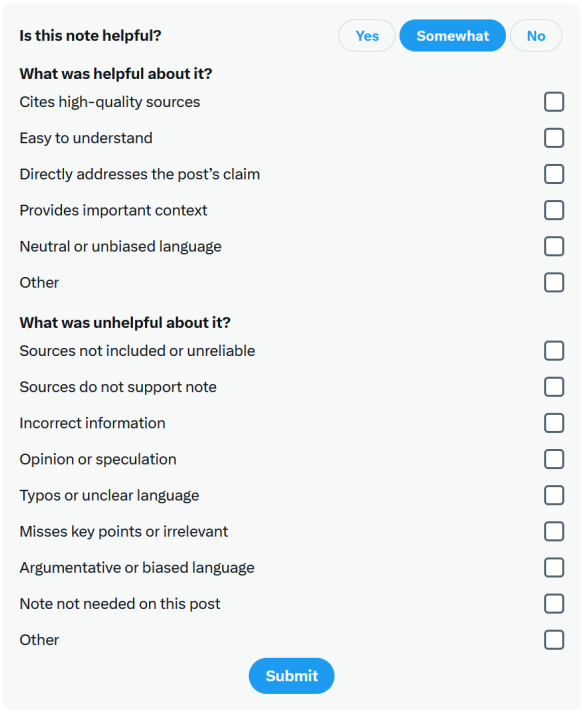


(b) Ratings per users or notes scaled down to disregard outliers.

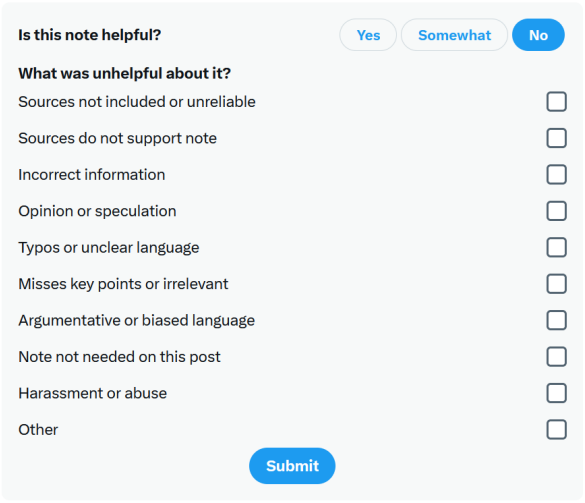
Fig. 3: The median of ratings on notes is 27 ratings. The 25 and 75 percentile are at 8 and 81 ratings. The median of the ratings per contributor is 33 ratings. The 25 and 75 percentile are at 9 and 103 ratings.

Further analysis of the Community Notes data shows that around 16% of the notes have less than five rating, meaning they are not disregarded in the note scoring process and receive a “needs more ratings” status.

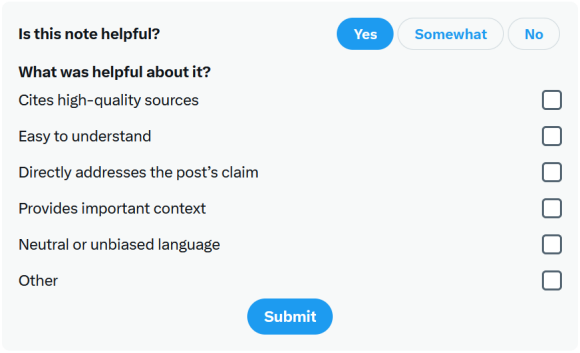
Note Tags



(a) Tag options for the "somewhat helpful" rating



(b) Tag options for the "not helpful" rating



(c) Tag options for the "helpful" rating

Fig. 4: The tags each contributor can assign to a note when they rate it either "helpful", "somewhat helpful" or "not helpful". The screenshots were obtained from the Community Notes user interface on the 23.07.2025.

Effect Network

Description of the System Variables

 # misleading posts without a note quantifies the number of misleading posts that did not receive a note.

Ratio of notes where raters with different factor vectors agree on helpfulness describes the number of raters with different factor vectors that agree on the helpfulness of a note compared to the overall ratings on the note.

Ratio of notes rated according to own viewpoint determines the number of notes that are rated according to the viewpoint of the rater compared to the overall number of ratings.

Ratio of notes evaluated according to their actual helpfulness determines the number of notes that are rated according to actual helpfulness of the note compared to the overall number of ratings.

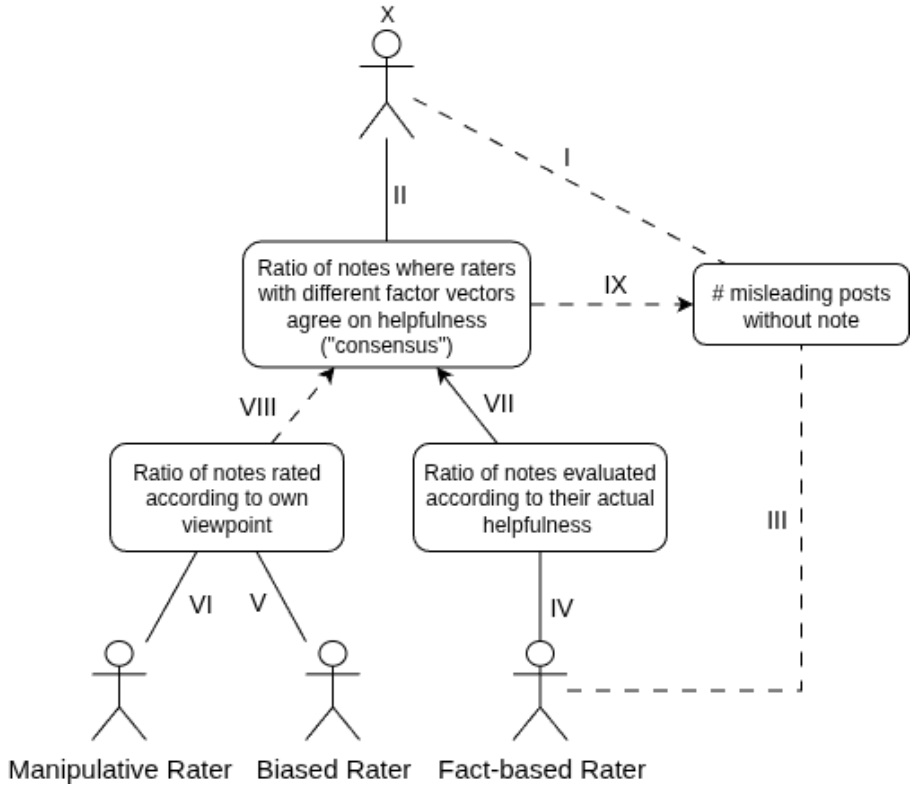


Fig. 5: The effect network shows the process of how the rating behavior impacts the number of notes that are shown on misleading posts.

Description of the Edges

Edge I represents the previously described motivations of X to achieve a high level of agreement among raters with different factor vectors on the helpfulness of notes.

Edge II represents the previously described motivations of X to reduce the total number of misleading posts without a note.

Edge III displays the motivation to reduce the number of misleading posts that do not receive a note.

Edge IV represents the motivation to rate as many notes as possible according to their actual helpfulness.

Edge V illustrates the subconscious motivation of the *Biased Raters* to rate as many notes that align with their beliefs as helpful as possible. Simultaneously, they strive to rate notes that do not correspond to their viewpoint as unhelpful.

Edge VI displays the motivation of the Manipulative Raters to rate notes according to an specific viewpoint with the goal to reinforce their own viewpoint or that of a certain agenda.

Edge VII An increase in the "ratio of notes evaluated according to their actual helpfulness" results in an increase of the "ratio of notes where raters with different factor vectors agree on the helpfulness" because then a high number of people agree that the note is either helpful or unhelpful. The note will finally receive the respective status.

Edge VIII An increase in the "ratio of the notes that are rated according to the viewpoint of the rater" results in a decrease of the "ratio of notes where raters with different factor vectors agree on the helpfulness". This assumption is supported by the fact that contemporary societies are highly polarized (Barber et al., 2015; Carothers and O'Donohue, 2019; Herold et al., 2024; Pew Research Center, 2014). As a result, when the ratings are primarily based on their alignment with the raters' perspectives, there are fewer notes where raters with different factor vectors agree on their helpfulness.

Edge IX displays that an increase in the number of notes where raters agree on the helpfulness leads to a decrease in the *number of misleading posts without a note*, as more notes will be displayed on X. It should be noted that the assumption here is that the displayed notes are indeed helpful.

References

Allen J, Arechar AA, Pennycook G, et al (2021) Scaling up fact-checking using the wisdom of crowds. Science Advances 7(36):eabf4393

Allen J, Martel C, Rand DG (2022) Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In: Proceedings of the 2022 CHI conference on human factors in computing systems, pp 1–19

Barber M, McCarty N, Mansbridge J, et al (2015) Causes and consequences of polarization. In: Political negotiation: A handbook, vol 37. Brookings Institution Press, Washington D.C., p 39–43

Carothers T, O'Donohue A (eds) (2019) Democracies Divided: The Global Challenge of Political Polarization. The Bookings Institution, Washington D.C.

Drolsbach CP, Solovev K, Pröllochs N (2024) Community notes increase trust in fact-checking on social media. PNAS Nexus 3(7)

Edwards K, Smith EE (1996) A disconfirmation bias in the evaluation of arguments. Journal of personality and social psychology 71(1):5–24

Godel W, Sanderson Z, Aslett K, et al (2021) Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. Journal of Online Trust and Safety 1(1)

Herold M, Joachim J, Otteni C, et al (2024) Polarization in europe: Results by country. URL https://forum-midem.de/2024/02/07/midem-study-polarization-in-europe-results-by-country/

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37

Lewandowsky S, Ecker UKH, Seifert CM, et al (2012) Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest 13(3):106–131

Martel C, Allen J, Pennycook G, et al (2024) Crowds can effectively identify misinformation at scale. Perspectives on Psychological Science 19(2):477–488

Olson DL, Delen D (2008) Advanced Data Mining Techniques, 1st edn. Berlin, Heidelberg

Pearl J, Glymour M, Jewell NP (2016) Causal inference in statistics: A primer. John Wiley & Sons, Hoboken

Pew Research Center (2014) Political polarization in the american public. URL https://www.pewresearch.org/wp-content/uploads/sites/4/2014/06/6-12-2014-Political-Polarization-Release.pdf

Reuters Institute for the Study of Journalism (2023) Digital News Report 2023. URL https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Schwarz N, Newman E, Leach W (2016) Making the Truth Stick & the Myths Fade: Lessons from Cognitive Psychology. Behavioral Science & Policy 2(1):85–95

Wojcik S, Hilgard S, Judd N, et al (2022) Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. arXiv:2210.15723

X (2024) Community notes: a collaborative way to add helpful context to posts and keep people better informed. URL https://communitynotes.x.com/guide/

Zweig KA, Krafft TD, Klingel A, et al (2021) Sozioinformatik : ein neuer Blick auf Informatik und Gesellschaft. Carl Hanser Verlag, München

Speech Act Gaps and Counter-Speech

Isaac Taylor[1]

[1] Department of Philosophy, Stockholm University, Sweden.

Corresponding author. E-mail: isaac.taylor@philosophy.su.se

Abstract

Speech produced by artificially intelligent systems is likely to become more ubiquitous in the future. This article examines potential costs of this trend, and explores ways of avoiding them. It is suggested that artificial speech may generate a "speech act gap": an apparent speech act that cannot be attributed to anyone. This may be problematic for various reasons, but one which is set out in the article is that speech act gaps will render counter-speech — forms of speech aimed at nullifying the negative effects of dangerous speech — less effective. If we are to avoid the negative effects that may result from dangerous artificial speech, alternative responses may be needed.

Keywords

Artificial Intelligence, Counter-Speech, Responsibility Gap, Speech Acts.

1 Introduction

Someone yells "Fire!" in a crowded theatre. The consequences of this might be quite significant. It can change the beliefs of the audience and cause a stampede, putting many people at risk of injury. This is because the shout of "Fire!" is not mere speech; it is also what philosophers of language call a "speech act". It is a way of doing things with words (Austin, 1962). Speech acts, because of their wider effects on the mental states and behavior of listeners can cause great harm, especially when their audience is expanded through new media. Hate speech can undermine the assurance and security of vulnerable groups (Waldron, 2012), and inducing false beliefs by spreading misinformation can prevent political communities from governing themselves in an autonomous manner.

How should we respond to these types of "dangerous speech" (Howard, 2019), as they can be called? Societies have developed a number of regulatory options. We might, for instance, simply ban it. But many have thought that a better – more effective and ethically acceptable – response is more speech. Instances of "counter-speech", speech acts which respond to other speech acts, might in one way or another neuter the problematic effects of dangerous speech. One simple form of counter-speech in response to misinformation, for instance, is to provide the correct information, backed up by evidence. The hope of many liberals is that the truth will win out in the marketplace of ideas, and that silencing falsehoods is unnecessary and unproductive (Mill, 1977).

In the future, more and more of the speech we encounter may be artificial: produced not by humans but by artificially intelligent systems. We are by now familiar with the use of bots on social media, which can post, respond, and interact with users (as well as other bots). Recent developments in large language models suggests that the sorts of text that can be produced will become more sophisticated. Despite these advances, it is likely that some of the artificial speech produced will be dangerous. Large language models have been known to "hallucinate" and produce false information (for example, misattributed quotes). And since AI systems in general are trained on masses of human data, the speech produced may reproduce biases and found in that data (and perhaps, on occasion, even generate something that looks like hate speech). On the other hand, some actors have put AI systems to use for malicious reasons - ranging from providing misinformation about vaccine effectiveness to voicing opposition to government policy (thereby inflating apparent discontent). We may be hopeful that our tried-and-tested practices of counter-speech can have a similar beneficial effect here. In this paper, however, I provide some reasons for doubt on this front.

Section 2 provides a theoretical model of what is happening when artificial speech is produced that resembles human speech. I characterize the situation as one of a "speech act gap": an apparent speech act without a corresponding actor. The speech act gap is morally significant, I claim, because its existence raises the possibility that some of the ways in which we respond to dangerous speech will be ineffective. I outline a collection of responses – known collectively as "counter-speech" – in section 3, focusing on one known as "blocking". Section 4 demonstrates how some of the benefits we may hope to achieve through counter-speech may not be attainable when it is in response to artificial speech which is not attributable to anyone. In particular, many of the sanctioning functions of counter-speech will not be achieved. I conclude by considering the wider lessons from this case for our regulation of our increasingly technologically-mediated social practices more broadly.

A note on the methodology used is in order. This is primarily a work of normative philosophy. While it relies on a number of (more or less) empirical claims (that AI systems do not have intentions, that harms conducted without an appropriate object of retribution will lead to undesirable outcomes, and so on), these are not defended in any great detail. They are, for the most part assumed for the purposes of examining the normative upshot. More work is almost certainly needed to address these sorts of questions – especially the ones about how we should understand what AI systems are doing. Nonetheless, the claims are ones that have been defended in the literature, and might furthermore be plausible on the face of it. My intention here is to draw out the implications of these claims for how we should organize our social practices.

2 Speech Act Gaps

How are we to interpret instances of artificial speech? While they might often look like speech acts, some authors have argued that this impression is misleading. These are not speech acts, it is claimed, because of the lack of intention behind them (Gubelmann, 2024), (Williams and Bayne, 2024). It does not look like AI systems can have intentions, understood as practical orientations towards specific goals.

While a human shouting fire in a theatre will generally do so with a specific goal in mind (causing panic, for instance), when AI systems generate text, they cannot be said to have any goal (which is not to say that their programmers had no goals in creating them in the first place). They are more akin to someone with a rare form of Tourette's syndrome who compulsively shouts "fire" in crowded spaces without any desire to do so. Large language models have thus been described as "stochastic parrots", mimicking human speech but never successfully performing the actions that speaking as a human at least sometimes involves (Bender et al., 2021). As indicated above, more work needs to be done on the internal representative capacities of AI systems, but I will assume for the moment that this picture is correct.

Why do intentions matter for the attribution of speech acts? More generally, I suggest, we want to distinguish intentional acts from non-acts because we would judge the person behind them differently. People who perform intentional acts are accountable for what they do. We can properly praise or blame them depending on the deontic status of the act performed. In contrast, if people do things unintentionally, it would often be improper, and unfair, to apply social sanctions like praise or blame, for those doings. If we learned that a given fire-shouter in a theatre has the form of Tourette's syndrome imagined above, any anger that we would otherwise have for them should dissipate. It is useful, for our practices of holding each other responsible, to have a distinction between acts and non-acts and, more specifically, speech acts and other forms of sounds that we can make. It seems, then, that artificial speech may often generate what I call a "speech act gap": an apparent speech act without an actor that it can be attributed to.

Even if we accept that artificial speech cannot be a speech act, it might nonetheless have some of the same effects. Artificial speech can be remarkably human-like. And, as noted above, it may be incorrect, biased (in the normative, and not statistical, sense of the term), or even derisory of individuals and groups.

To see what is going on here more clearly, we might draw on a distinction made by Austin between the illocutory and perlocutionary parts of a speech act (Austin, 1962). The illocutory part refers to the intended function of a particular utterence: whether it is supposed to assert something, request something, apologize to someone, and so on. An individual who shouts "Fire!" might intend to warn others in the immediate vicinity, for instance. The perlocutionary part of the speech act refers to the actual effects on listeners: a change in beliefs, a certain action, and so on. Those who hear the cry of "Fire!" might be sent into a panic and attempt to run. Another way of putting the point here is that artificial speech cannot have an illocutionary element (owing to the lack of intention in the systems that produce them), but it can nonetheless produce the same perlocutionary effects, good and bad. If we think that something needs to be done in response to dangerous speech acts made by humans, owing to the perlocutionary effects (believing false information, for example), then it stands to reason that we need to consider what should be done in response to dangerous artificial speech.

At this point, it is worth drawing attention to parallels between the speech act gap and another concept found in the field of AI ethics. It is often claimed that the use of AI systems to make high-stakes decisions risks generating problematic "responsibility gaps": situations in which nobody is responsible for that decision (Mathias, 2004), (Sparrow, 2007). This is especially thought to be true with respect to autonomous weapons systems, which would be able to select and engage targets in armed conflicts without direct control by human actors. It is thought by some that the use of such weapons would render nobody responsible for the deaths caused. But responsibility gaps might plausibly arise whenever an AI system is used to make or assist with a high-stakes decision. This could be a decision about where a car should swerve in an accident scenario, how long a given defendant should spend in jail, or which course of treatment a patient should receive. While some authors claim that, despite appearances, we can locate appropriate loci of responsibility, perhaps with the help of organizational or technological interventions, limits have been noted to these strategies (Taylor, 2024), (Taylor, 2025a).

It may be asked why a lack of responsibility would be problematic. Various benefits of maintaining responsibility have been noted in the literature. I mention two here. These may not be the most significant but, as will become clear in due course, they are the most relevant to the present discussion.

First, maintaining responsibility might be important because it will ensure that unjust harms are minimized (Baum et al., 2022), (Sparrow, 2007). If people are aware that they are responsible for certain decisions, they will take care in ensuring that the decisions in question are morally acceptable. Their motives for this might vary: a moral motivation to not be complicit in injustice might be one, but fear of social sanctions might motivate them just as well. In contrast, if nobody is responsible for

a decision, nobody will be potentially subject to these sorts of pressures, and consequently they may lack the incentive to ensure that unjust harms are minimized. Call this the "instrumental benefit" of responsibility.

Second, responsibility might be important to ensure an appropriate object of retributive blame and punishment when harmful decisions are reached. John Danaher has argued that responsibility gaps may give rise to "retribution gaps", whereby a deep-seated desire to respond to (apparent) injustice is met with no appropriate object of retribution (Danaher, 2016). It is widely thought that blame and punishment are only appropriately directed at those who are guilty of wrongdoing and yet, if there is a responsibility gap, there are no such agents. Danaher suggests that retribution gaps may, in turn, lead to scapegoating or a loss of faith in the rule of law. When someone is responsible, on the other hand, sanctions can be applied without wronging those they are applied to. Call this the "retributive benefit" of responsibility.

Two similar benefits might be thought to attach to speech retaining its status as a speech act. There might, first, be an instrumental benefit of speech acts: if speech is attributable to intentional actors, there will be incentives to be careful about what is said, lest one's conscience or other listeners disapprove. In particular, a speaker may try to avoid engaging in dangerous speech to avoid these internal and external sanctions. And there might, second, be a retributive benefit of speech acts: if a piece of dangerous speech has an actor, they will be the appropriate object of retributive punishment or blame for it. Such sanctioning may be important to restore the faith of those harmed that their status as a source of valid moral claims is taken seriously by their community. Condemnation, for instance, might be important in restoring the self-respect of those who are the victims of hate speech. It demonstrates that the sentiments expressed are not more widely-held (Waldron, 2012).

These benefits may be lost if there is a speech act gap: there will be no incentives to avoid dangerous speech, and no appropriate object of retribution when dangerous speech takes place. To see this more clearly, it is worth examining the means of responding to harmful acts that must be on hand to achieve both instrumental benefits and retributive benefits. In the case of decision-making, this will primarily be the application of sanctions. In the case of speech acts, this will often be through counter-speech.

3 Counter-Speech

Counter-speech refers to a variety of techniques that seek to minimize the harms of dangerous speech (Cepollaro et al., 2022). As noted above, this might simply involve correcting misinformation. Other examples include openly condemning or providing alternative narratives. Here, however, I focus on a more complex form of counter-speech that has been of interest to philosophers of language in recent years: blocking. I focus on this because, as we will see, the preconditions to blocking might be especially unlikely to be met if we are dealing with artificial dangerous speech.

Dangerous speech can take many subtle routes to causing harm. Consider the following proposition:

"Martin Luther King Jr. was an eminent scholar despite being black."

This statement involves what Rae Langton calls a "back-door speech act". Although not explicitly stated, the word "despite" suggests that black people tend to lack the capacities to be good scholars. According to Langton, if such back-door speech acts go unchallenged, the propositions they imply become part of the common ground – what is taken for granted – in subsequent discussions (Langton, 2017).

It may be important, then, to challenge back-door speech acts like these, removing them from the common ground. Langton argues that one can "block" a backdoor speech act by questioning the undefended presuppositions that they involve (Langton, 2017). In our example, all this might require is someone speaking up and asking why the word "despite" is being used. The original speaker now has two options: they can either defend the view that motivated the back-door speech act (which may or may not be successful) or they can back down. Blocking thus has the potential to 'undo' the negative effects of backdoor speech acts, and remove their implicit propositions from the common ground.

Suppose that the sentence discussed above was put forward, not by a human, but an AI system. Depending on the context, it might have the same negative effects. If multiple instances of artificial speech all involve the implication that black scholars tend not to be doing valuable work, this may end up contributing to the creation of a society where this is taken for granted. Can we nullify those effects in the same way as we might hope to do with certain forms of human speech, namely, through counter-speech?

One initial worry we might have is that, given the sheer quantity of artificial speech we might see in the near future, responding to each and every instance of dangerous artificial speech would be infeasible. Bots could produce multiple online posts in a fraction of the time it takes a human to construct one. However, it should be noted that we could, in turn, rely on AI systems to engage in counter-speech, and ensure that it is done equally quickly. We already see this in the correction of misinformation through automated fact-checking. Is there any reason to think that more complex forms of counter-speech, like blocking, could not also be done artificially?

AI systems might be quite effective in identifying dangerous speech online, even subtle forms involving back-door speech acts. I fed in the sentence above into a well-known large language model, and found that it was able to identify the problematic implication. We might speculate that this was because it takes the form of many other examples of backdoor speech acts found in the literature in the political philosophy of language (even if this precise one, to my knowledge, has not been mentioned in this literature before). Perhaps there will be some formulations that we are not so familiar with which AI systems would have a harder time identifying.

It is worth distinguishing here between what we might call automated and autonomous AI systems. The distinction, or at least one way of making the distinction, rests on how much their initial programming specifies their behavior (Sartor and Omicini, 2016). More automated systems are given less leeway in advance about what they will do – programmers tend to have more of a hand in specifying how they will behave in response to certain inputs. If we are thinking about bots, for instance, the programmers might program a system to display a specific message when they encounter a post of a certain type online. More autonomous systems, on the other hand, will have no such precise specifications. Bots incorporating large language models, for instance, develop novel responses to text input which have not been specified by anyone in advance.

If we are dealing with autonomous systems, I posit, there is a greater possibility of novel (apparent) backdoor speech acts. Any merely automated systems that were designed to put forward problematic speech would have to rely on existing known formulations. When we are dealing with autonomous systems, in contrast, there may be apparent backdoor speech acts generated as a result of biases in the data (for instance, a lack of recognition of black scholars' work). There is no reason to think that these would take the form that we are familiar with. Consequently, we may worry that the increasing need for artificial counter-speech, brought about as a result of the increasing ubiquity of artificial speech, may undermine our practices of blocking. AI systems may not have the necessary capacities to effectively identify backdoor speech acts that should be challenged.

Whether or not this problem extends to other forms of counter-speech may be questioned. AI systems may be quite effective in correcting false information. However, online speech may be misleading in different ways than this: by missing out crucial context accompanying information, for instance. Could an AI system identify the lack of context and provide it? I set this question aside, since I think that there may be a less contingent problem with responding to dangerous aritifical speech, whether this response needs to be done artificially or can be left to humans. This relates to the wider benefits that our practices of counter-speech might normally have, but which may dissipate if counter-speech responds to artificial speech.

4 The Wider Benefits of Counter-speech

We noted above that avoiding a speech act gap may ensure both an instrumental benefit and a retributive benefit. Note that these benefits may require at least the possibility that counter-speech will follow speech acts. Bad-faith individuals may only be incentivized to avoid dangerous speech if they think that their misinformation will be publicly corrected, their hate speech will be challenged, or their back-door speech acts will be blocked. Such counter-speech may not only "undo" the harms of dangerous speech. It may also provide a social sanction, which some may not want to risk being subject to: the shame of being challenged may be sufficient to disincentivize dangerous speech. This sanctioning role of counter-speech also explains how retribution can be carried out in some form when we can attribute speech acts to individuals. Counter-speech can (at least implicitly) involve blaming people for what they have said.

It is worth noting one initially plausible, but ultimately unconvincing, reason why the sanctioning role of counter-speech might not be possible when we are dealing with artificial speech. As I said in the previous section, owing to the sheer quantity of artificial speech we may see in the future, we may need to rely on AI systems to conduct counter-speech as well. It may, then, be questioned whether speech by an AI system can involve any sanctioning at all. After all, the sorts of negative reactive attitudes that

sanctioning may seem to presuppose are not possible in existing or near-future AI systems. Nonetheless, I suggest, sanctioning can still occur – and the negative reactive attitudes required for it can be traced back to humans. This would require either that humans maintain control over the system (and it may matter *who* is in control (Taylor, 2023)) or else if the system can be understood as acting in the name of humans (which it might do if its actions have a suitably robust connection to their values, reasons, and the like (Taylor, 2025b)).

The real barrier to the successful sanctioning of artificial speech comes not from the nature of the counterspeaker but rather from the nature of the primary speaker. Consider what would happen if a piece of dangerous speech, say a problematic (apparent) back-door speech act, was put forward by an AI system. There may indeed be counter-speech – either by another human or by an AI system here (although, as we saw in the last section, whether it will occur in practice may not be guaranteed). But, since the apparent speech act itself cannot be attributed to any human, any incentives to avoid dangerous speech that are generated by the possibility of counter-speech may not be present. We might, of course, hold the designers of the system liable to pay a fine or compensation for acts of dangerous speech. But the social sanctioning that is often carried out by counter-speech would not be present in such a case, as any counter-speech could not be properly understood directed to those who develop or deploy autonomous systems. If someone is to be effectively criticized for a speech act, it must be one that can be attributed to them. The systems themselves are not the sort of entity capable of feeling the sting of social criticism, and thereby adjusting their behavior in response.

It is important to note that there are some (potentially desirable) incentives that might be created through counter-speech to artificial speech. If, for instance, a backdoor speech act by an autonomous bot is challenged, this may signal to others that this is not the sort of speech that will be tolerated and the presuppositions in it will not be taken into the common ground without question. Observers may, therefore, decide to adjust their behavior accordingly. But this will do little to disincentivize further dangerous speech by other AI systems (or indeed, the system in question on this occasion), since those systems themselves would not respond to sanctions in the same way.

For similar reasons, a retribution gap may occur in such cases as well. When the systems that are generating dangerous speech are sufficiently autonomous, and thus the speech cannot be predicted by those behind the system itself, it will be inappropriate to blame the latter. The standard conditions on blameworthiness – which require both knowledge about and control over the action for which one is being blamed for – are not met. Consequently, there is no appropriate object of blame. If an AI system produces something that looks like a problematic backdoor speech act, for instance, while we might be able to question the presuppositions involved, doing so could not properly entail blame.

This suggestion might be too quick, however. We might want to distinguish here between what we might call deceptive and non-deceptive systems. Non-deceptive systems would be systems that are presented as AI systems. Deceptive systems, on the other hand, produce outputs that are presented as human speech.

If we are dealing with deceptive systems, counter-speech may be enough to fill the retribution gap. Since onlookers will assume that retributive sanctioning has been applied to an entity that deserves it (namely, a human speaker), whether or not actual retribution takes place seems somewhat beside the point. People would *feel* that retribution has taken place, and this may be enough to avoid the negative effects that Danaher identifies (scapegoating and a loss of faith in the rule of law). The problem, however, will arise with non-deceptive systems. Any response to the outputs of these systems, in whatever form it takes, may well be unsatisfying to onlookers. Since the AI system itself is not a being capable of experiencing the sting of social sanctioning, there may still be a retribution gap after counter-speech has been undertaken.

It might be suggested that there is no problem here. After all, individuals will know that non-deceptive systems are not really actors, and they might thus not desire retribution when these systems produce dangerous speech. Something similar might be true with respect to other sorts of AI systems: if we know that a harm has occurred as a result of behavior by AI systems, any urge we have to punish the perpetrator in some way may dissipate as we realize that, strictly speaking, there is no perpetrator.

Danaher seems unconvinced that things will be that easy, however. One of the premises in his argument is that

> If an agent is causally responsible for a morally harmful outcome, people will look to attach retributive blame to that agent (or to some other agent who is deemed to have responsibility for that agent). (Danaher, 2016)

Since robots might be understood as agents (Floridi, 2025), any harm caused by them will, if Danaher is correct, be met with this desire on the part of humans. It may be quite a deep-seated reaction, and Danaher provides evidence that it is widespread (Danaher, 2016). Of course, we should do something about this if it causes the negative consequences identified (Kraaijeveld, 2020). But that might be easier said than done. Similarly, if we knew that AI systems, and not intentional agents, were producing dangerous speech, it would be ideal if we reduced or eliminated our retributive urges. But that might not be so easy either, given that these reactions are hardwired into us.

I conclude that some of the positive effects of counter-speech may not be present when it is undertaken in response to artificial speech. In particular, the social sanctioning – which is valuable for a variety of purposes – may not be attainable. Whether or not this problem arises in practice, however, depends on the nature of the AI system.

5 Conclusion

I have argued that the liberal hope that more speech is often the best response to dangerous speech – perhaps naïve even before the "fourth revolution" where our online and offline lives have become intertwined (Floridi, 2014) – may need to be re-thought with the increasing ubiquity of artificial speech. The broader lesson to draw from all of this is the following: interventions that were designed for regulating human speech in desirable ways may not have the same effects when drawn on in regulating artificial speech.

In closing, it is worth exploring – in an unavoidably provisional fashion – what other strategies we have for responding to dangerous artificial speech. The liberal interest in counter-speech, in part, grew out of a worry about placing restrictions on freedom of speech in the first place. More so of the worry about placing restrictions on freedom of thought. Yet, while restricting the speech and thought of humans might be viewed as especially problematic, existing AI systems do not have the sorts of interests that warrant rights (Basl and Bowen, 2020). And it is implausible to suggest that restricting the "speech" (read: outputs) and "thought" (read: internal processing) of these systems restricts the freedom of speech and thought of their users or developers. After all, of these speech acts cannot be attributed to anyone (as the speech act gap hypothesis holds), they cannot be thought of as an extension of the expression of the humans behind them.

All this suggests that a different approach, where restrictions are placed on the development and use of AI systems that produce artificial speech might be permissible from a liberal perspective. We might, for instance, silence bots that are detected as producing dangerous speech by removing their content. How we can distinguish bots from real humans, however, is a practical question that would need to be answered if this is to be pursued. Alternatively, we might restrict the freedom of "thought" of bots by requiring that they are programmed in ways that avoids reproducing biases and stereotypes. As AI technology becomes more accessible to different actors, however, the effectiveness of such legislation may be doubtful. In addition, such an approach would need to be implemented carefully in order to avoid stifling innovation and restricting too much content. Finally, it is worth noting that there may be reasons for avoiding censorship that are not exhausted by consideration of the rights of those censored. Such reasons may extend to censorship of artificial speech.

Given the difficulties of this sort of approach we might, taking a cue from the literature on the responsibility gap (Nyholm, 2018), attempt to bridge the speech act gap by attributing dangerous speech acts to the collaborative agency involving both AI systems and the humans behind them. Whatever the merits of this approach for dealing with the responsibility gap, there seem to be additional difficulties when we move to the speech act gap. In particular, while we are familiar with ideas of collective agency, the idea of collective speech seems to be more of a stretch. A workable model of this would need to be developed in order to demonstrate that this strategy might have promise.

Acknowledgements. I am very grateful to the participants at the Paris Conference on AI and Digital Ethics in 2025, as well as the reviewers of this journal, for feedback. I would also like to thank Anandi Hattiangadi for discussions that sparked my interest in these issues.

Declarations

Conflict of Interest: The author declares no conflict of interest.

References

Austin JL (1962) How to do Things with Words. Clarendon Press, Oxford

Basl J, Bowen J (2020) Ai as a moral right-holder'. In: Dubber M, Pasquale F, Das S (eds) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford, p 289–306

Baum K, Mantel S, Schmidt E, et al (2022) From responsibility to reason-giving explainable artificial intelligence. Philos Technol 35. 10.1007/s13347-022-00510-w

Bender EM, Gebru T, McMillan-Major A, et al (2021) On the dangers of stochastic parrots: can language models be too big? FAccT '21 10.1145/3442188.3445922

Cepollaro B, Lepoutre M, Simpson RM (2022) Counterspeech. Philos Compass 18(1). 10.1111/phc3.12890

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18(4):299–309. 10.1007/s10676-016-9403-3

Floridi L (2014) The 4th Revolution: How the Infosphere is Reshaping Human Reality. Oxford University Press, Oxford

Floridi L (2025) Ai as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. Philos Technol 38. 10.1007/s13347-025-00858-9

Gubelmann R (2024) Large language models, agency, and why speech acts are beyond them (for now) – a kantian-cum-pragmatist case. Philos Technol 37. 10.1007/s13347-024-00696-1

Howard J (2019) Dangerous speech. Philos Pub Aff 47(2):208–254. 10.1111/papa.12145

Kraaijeveld SR (2020) Debunking (the) retribution (gap). Sci Eng Ethics 26(3):1315–1328. 10.1007/s11948-019-00148-6

Langton R (2017) Blocking as counter-speech. In: Harris D, Fogal D, Moss M (eds) New Work on Speech Acts. Oxford University Press, Oxford, p 144–164

Mathias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183. 10.1007/s10676-004-3422-1

Mill JS (1977) On liberty. In: Robson JM (ed) Collected Works of John Stuard Mill. Toronto University Press, Toronto, p 213–310

Nyholm S (2018) Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. Sci Eng Ethics 24(1):1201–1219. 10.1007/s11948-017-9943-x

Sartor G, Omicini A (2016) The autonomy of technological systems and responsibilities for their use. In: Bhuta N, Beck S, Geib, et al (eds) Autonomous Weapons Systems: Law, Ethics, Policy. Cambridge University Press, Cambridge, p 39–74

Sparrow R (2007) Killer robots. J App Philos 24(1):62–77. 10.1111/j.1468-5930.2007.00346.x

Taylor I (2023) Justice by algorithm: The limits of ai in criminal sentencing. Crim Justice Ethics 42(3):193–213. 10.1080/0731129X.2023.2275967

Taylor I (2024) Collective responsibility and artificial intelligence. Philos Technol 37. 10.1007/s13347-024-00718-y

Taylor I (2025a) Is explainable ai responsible ai? AI Soc 40:1695–1704. 10.1007/s00146-024-01939-7

Taylor I (2025b) Representative robots: Can ai systems act in our name? Free Equal 1(1):298–322. 10.16995/fe.18626

Waldron J (2012) The Harm in Hate Speech. Oxford University Press, Oxford

Williams I, Bayne T (2024) Chatting with bots: Ai, speech acts, and the edge of assertion. Inquiry 0.1080/0020174X.2024.2434874

Accountability and Attribution in AI-Generated Content Authentication: Lessons from China's AI Content Labelling Mandate

Wenlong Li[1]

[1] Independent Researcher.

Corresponding author. E-mail: fettes.lee@gmail.com

Abstract

This paper interrogates how China's pioneering AI content labelling mandate responds to urgent, transnational concerns of provenance and authentication concerning AI-powered content generation and dissemination. Effective 1 September 2025, this unprecedented regulatory framework compels AI providers, deployers as well as distribution platforms to implement a bifurcated schema of explicit and implicit content labelling, thereby prioritising unambiguous provenance and seamless traceability across the entire lifecycle of AI-generated artefacts.

It is observed that China's regime is aggressively prescriptive, hierarchical and deterrent, positioning itself as a laboratory for rigorous attribution. Essentially, China's approach does not merely mitigate digital deception domestically, but also reconfigures international regulatory paradigms by pairing technical mandates with a robust organisational extensive architecture for accountability. It actively reconfigures platform and provider incentives at scale, creating a testbed for enforcing provenance solutions.

The resonance of these measures is unmistakably global: parallel developments across the world (e.g., the EU's AI Act and emergent statutes in some US states), some of which have seemingly influenced China's development, echo its commitment to provenance, and China's action both reflects and shapes the emergent consensus. Rather than advancing prescriptive solutions or engaging with comparative analysis, this short piece offers a measured examination of China's regulatory experiment as a vantage point for broader reflection. It illuminates the underlying questions and emerging dilemmas in the discourse around AI authentication and attribution and leverages China's leading steps to provoke critical thought about the possibilities and limitations of governance strategies in a rapidly evolving digital environment.

Keywords

China, AI-Generated Content, AI Supply Chain, Content Labelling, AI Regulation.

1 Introduction

The exponential rise of AI-generated content has precipitated a moment of reckoning for global regulatory and ethical discourse. As generative model now generates text, image, video and audio that defy effortless human discernment, the prospect of a reality saturated with indistinguishable synthetic media has provoked urgent demands for both robust content authentication and sophisticated governance architectures. Recent episodes – including high-profile deepfake political interventions that have reshaped electoral narratives(Kapoor and Narayanan, 2024; Łabuz and Nehring, 2024) or the viral dissemination of fabricated news orchestrated by advanced AI (Adami, 2024) – have accentuated the need for mechanisms that ensure transparency and accountability in an increasingly synthetic information environment.

The challenge of AI content authentication is further magnified what (AI) ethicists term the 'many hands' problem: as AI-generated content initiates within sophisticated technical pipelines and subsequently traverses a myriad of hands – including developers, platform intermediaries, automated distributors, and global user networks – attribution and accountability are dispersed to the point of opacity(Nissenbaum, 1996). The literature on the AI supply chain stresses the fractal complexity of these pathways, cautioning that authentication must grapple not only with technical provenance, but also with the unpredictable and collective agency embedded in a polycentric, trans-platform content ecosystem(Cobbe et al., 2023a). Crucially, the challenge of content authentication extends beyond the purview of direct supply-side actors. As AI-augmented content migrates through social networking sites, news aggregators, messaging services, audiovisual streaming platforms, and the expansive terrain of user-generated communities, the dissemination mechanism become radically polycentric. This diffusion does not merely obscure provenance but multiples vectors for manipulation and abuse, thereby intensifying the risks posed by harmful or deceptive content.

Imagine that a wellness influencer launches a series of AI-generated explainer videos touting a new dietary supplement. As the content spreads, marketing agencies splice in fabricated user testimonials and doctor endorsements produced by text-to-video AI systems. The videos then move downstream: affiliate marketers employ additional automated tools to dub the material into several languages and inject trending but wholly unfounded health claims tailored to different markets. On social media and messaging apps, virality takes over, leading to the remixing of the original videos into short clips and meme, many stripped of context or transformed to highlight unproven cures. Influencers and micro-celebrities join the trend, further amplifying exposure with personalised stories, while e-commerce vendors automate links to dubious supplement sellers beneath each post. This scenario isn't far from reality and begs a host of persistent questions: how should liability be assigned across the tangled web of originators, curators, distributors and users? What regulatory and technical mechanisms might reliably reconstitute provenance when the chain is so easily broken? Can any system of accountability remain effective in a landscape defined by constant remixing, opacity, and the expanding capacity of both people and artificial agents?

The intricacy of this scenario, marked by relentless remixing and indeterminate lines of agency, starkly reveals the profound uncertainties and unresolved questions that contemporary governance regime must now confront. The European Union's Artificial Intelligence Act (AIA) stands as the first comprehensive attempt to address these concerns, introducing binding obligations such as the disclosure of synthetic media and machine-readable standardised labels for AI-generated content (EU, 2024). Across the Atlantic, the United States presents a markedly fragmented landscape featuring a mosaic of state-level statutes (notably in California(California, 2024) , New York (N.Y., 2024) (US, 2025) and Washington(Washington, 2025) among a few others[63]) still under deliberation, some ill-fated federal bills(US, 2023b) [64], and voluntary industry commitments encouraged by the then Biden administration(US, 2023a)[65] In contrast, China has instituted perhaps the most far-reaching and rapidly

[63]Other developments at state levels include Wisconsin (Senate Bill 644, mandating disclosures in political ads if they contain AI-generated synthetic media), Tennessee (Senate Bill 2431, requires disclosure for AI-generated content in Tennessee using another's likeness or falsely attributed authorship), Oklahoma (Senate Bill 746, mandating disclosure of AI-generated content in political ads, requiring a statement indicating AI involvement in creation). Notably, House Bill 2094 from Virginia that would have defined certain types of AI as "high risk" and placed transparency and reporting requirements on their usage was vetoed by their governor. see Andrews (2025)

[64]The Federal Communications Commission (FCC) announced a proposal on 25 July 2024 on Disclosure and Transparency of Artificial Intelligence-Generated Content in Political Advertisements, which would require broadcast stations to disclose AI-generated content through on-air announcements, defining such content as images, audio or video created using computational technology. see Commission (2024)

[65]Following the Biden administration's soft-handed regulatory approach to AI—including Executive Order 14110, AI policies in the United States has fundamentally shifted since Trump's return to office; the Trump administration rapidly rescinded key Biden-era directives and, under its new AI Action Plan, is now pursuing an aggressive program of deregulation. See House (2025).

implemented interventions to date, enacting mandatory requirements for the tracing and labelling of AI-generated content. Despite its historical and global significance as the first of its kind, developments in China remain under examination within much of western policy discourse. It is precisely this overlooked Chinese trajectory, with distinct legal architecture and implications for global governance, that forms the focus of this paper. To be specific, this short piece seeks to clarify China's rapidly evolving statutory and administrative regime of AI content authentication. It does not attempt to perform a comprehensive comparative analysis, and any reference to transnational borrowing or divergence serves to situate China's trajectory within the increasingly interconnected global evolution of AI governance.

It is argued that China's AI content labelling mandate stands as an unprecedented statutory response to the complex, polycentric reality of AI-generated content. In essence, it seek to proactively "stitch together" a continuous chain of accountability, binding service providers, distribution platforms and even end-user activities with an auditable compliance framework even as technical fragility persists. At the core of this regime is the transformation of "collective responsibility" that shifts the enforcement bottleneck from upstream providers to content distribution infrastructures, leveraging platform centrality in digital content flows as a regulatory fulcrum (Mühlhoff, 2025; Taylor, 2024). The stability and persistence of provenance metadata – central to China's regulatory architecture – remain vexed by technical and economic realities. Hence, its effectiveness will likely turn on the regime's ability to close technical loopholes, balance innovation with oversight, and coordinate standards both domestically and internationally. In practice, perfect compliance is improbable, making broader, layered approaches to accountability and risk mitigation essential. The Chinese mandates anticipate these weaknesses by criminalising removal or alteration of markers and prohibiting the distribution of tools for metadata erasure[66], but the technical arms race between robust provenance and circumvention tools is likely to persist, posing persistent threats to operational scalability. By channeling legal duties through a dense regulatory nexus – where platforms are accountable not only for detecting and flagging but also for reinforcing and tracing every instance of AI-generated content – the regime imposes significant compliance demands and operational costs, particularly for smaller developers or cross-border actors. The regime's success ultimately hinges less on perfect technical integrity than on the visibility, deterrence, and enforceability of its mandates. Large commercial providers and mainstream platforms are more incentivised to comply, but enforcement against smaller developers, open-source deployments, and gray-market actors remains a critical bottleneck.

This article is structured as follows. Following the introduction, Section II elucidates the conceptual and technical architecture of AI-generated content authentication, interrogating the technological complexities that render the tracing and substantiation of synthetic media particularly challenging. Section III pivots around China's trailblasing regulatory intervention in AI content labelling, offering a precise synopsis of the pertinent statutory and administrative obligations. Section IV critically assesses how its regime has grappled with the intricacies of provenance verification and authentication while exposing its blind spots, structural vulnerabilities, and regulatory lacunae. The final section situates these developments within the wider transnational and global regulatory landscape, drawing connections to ongoing international momentum and highlighting their prospective significance for the evolving terrain of AI governance.

2 AI-Generated Content Authentication: A Primer

Artificial intelligence-generated content authentication is undergoing a rapid metamorphosis, driven by the meteoric rise of synthetic media and an ever-more pressing need to safeguard digital trust. At its heart lies an intricate interplay of provenance, authenticity, and watermarking—concepts inseparably bound in their mission to combat digital deception. Content provenance emerges as the "digital paper trail" of our era: a meticulously curated ledger chronicling a digital asset's origin, creative process, and every intervening alteration (Werder et al., 2022). It weaves together details of authorship, chronology, locale, and the myriad transformations a piece of content endures as it traverses digital ecosystems. Provenance extends traditional notions of authenticity into the digital realm, ideally leveraging technological methods – such as cryptographic signatures and tamper-resistant metadata – to help users verify that certain content is genuine and unaltered (Longpre et al., 2024b; Pan et al., 2023). Authenticity, in this context, is the degree to which digital material can be reliably traced to its original source and verified as untampered. Provenance data underpin this authenticity, offering assurance against digital manipulation,

[66] Art 10, 13 MLAIGSC.

misinformation and the increasing sophistication of synthetic media generated AI (Bereskin, 2023; Feher, 2025; Schick, 2020). A case in point is the Coalition for Content Provenance and Authenticity (C2PA), which stands as an open, cryptographically assured standard for tracking digital content origins and edits. Its Content Credentials bind provenance data to content using digital signatures, and any change severs this cryptographic bond, making tempering immediately evident across platforms and throughout the content lifecycle. [67]

A distinction must be made among the various labels and markers deployed within authentication frameworks. Occasionally, the term "watermark" is indiscriminately applied to any conspicuous textual overlay—such as "DRAFT" stamps or "Not for Release" banners—that signal provisional status or restricted access (Council, 2024a). Yet, in the specialised context of digital authentication, watermarking takes on a more nuanced meaning. It signifies the deliberate embedding of subtle signals or patterns within digital content—crafted to be virtually invisible or inaudible to the unaided senses (Srinivasan, 2024). Unlike overt overlays or pictorial icons, these sophisticated marks are engineered to elude ordinary perception, surfacing only through the use of specialised algorithms or proprietary decryption keys. This paper embraces both visible and veiled modes as integral to contemporary authentication strategies and confines the term "watermarking" to its most precise sense: the surreptitious infusion of hidden, machine-detectable markers, intended exclusively to establish provenance and safeguard authenticity.
A multitude of techniques—none of which serve as a universal panacea—exist for content authentication. A time-honoured mainstay (and also the linchpin of China's approach as revealed later) is metadata: embedded digital records chronicling a file's origin, authorship, and processing history (Cheney et al., 2009; Deelman et al., 2009; Simmhan et al., 2005). Yet, relying on metadata is fraught with technical and organisational vulnerabilities (Hart and de Vries, 2017). Many mainstream social media, content-hosting, and file-sharing platforms routinely excise metadata (such as EXIF data or creator tags) when users upload or download material, citing privacy, file size and interoperability concerns.[68] Further, metadata can be manipulated or forged by adversaries and even well-intentioned users, thanks to the ready availability of metadata editing tools. While advanced metadata standards or cryptographically signed provenance records offer a promising remedy to achieve tamper resistance or evidence, their effectiveness hinges on seamless, consistent support across a chaotic digital landscape—otherwise, provenance can be irretrievably lost as content circulates between systems adhering to divergent standards or formats.

Advanced statistical watermarks present a sophisticated alternative to conventional metadata by weaving subtle markers directly into the very structure of digital outputs—manifesting, for instance, as unique pixel distributions in images or as distinctive token patterns within text (Srinivasan, 2024). These covert techniques, while ingenious, are not immune to tampering. Common manipulations such as format conversion (for example, switching from PNG to JPEG for images, or re-encoding textual data) can easily disrupt or completely obliterate the fragile statistical fingerprints that constitute these watermarks (Ren et al., 2024). Furthermore, adversaries with a vested interest in erasing provenance can deliberately employ strategies like image augmentation, paraphrasing, or noise injection—essentially randomising or mimicking watermark characteristics to obfuscate the content's origin or to spoof detection mechanisms(Xu et al., 2024). The absence of universal standards further compounds these vulnerabilities: many watermarking schemes are proprietary and tailored to specific models, so even robust watermarks may fail if every participant in a distribution chain is not equipped for their detection and validation. Open-source models further pose a structural challenge, as technical protections can be swiftly disabled by users with minimal sophistication(Srinivasan, 2024). Therefore, watermarking is thus adept at "raising the bar" for evaders, primarily within closed, commercial ecosystems, but cannot be seen as a fail-safe option.

[67] C2PA provides an open technical standard enabling digital media (such as images, video, audio, and documents) to carry cryptographically signed provenance metadata, known as Content Credentials. These credentials record who created the content, when, how, and with what tools, as well as subsequent edits or incorporated elements, all in a tamper-evident, verifiable structure. The standard does not act as DRM; rather, it promotes transparency regarding content origin, edit history, and chain of custody. Removing or altering these credentials is detectable because cryptographic hashes would fail to validate. C2PA is designed for compatibility with common metadata standards (like IPTC, XMP, EXIF) and can be used offline, with only a minimal increase in file size. The standard and its Trust List ensure that only credentials issued by certified authorities are considered trustworthy. For more, see the C2PA FAQ.: https://c2pa.org/faq/ (accessed 2025-07-31).

[68] The Social Media Sites Photo Metadata Test evaluated 15 top social media sites, and checked if embedded metadata was retained and displayed on upload to the sites or downloads of various version of the image. The results are displayed at www.embeddedmetadata.org/testresults. Only one social media site, Behance, received favourable results for retaining and displaying embedded data. A few systems retained embedded metadata but failed to use it when displaying metadata on the web site. Ten sites removed at least some metadata when images were downloaded to a desktop environment. https://iptc.org/news/many-social-media-sites-still-remove-image-rights-information-from-photos/, see also https://www.1854.photography/2013/03/study-exposes-social-media-sites-that-delete-photographs-metadata/ or https://blogs.loc.gov/thesignal/2013/04/social-media-networks-stripping-data-from-your-digital-photos/ See also (Irwin, 2024).

Two additional techniques—post-hoc detection and retrieval-based methods—expand the capabilities for authenticating AI-generated content. Post-hoc detection leverages machine learning to identify subtle statistical anomalies that suggest AI authorship. While such methods are appealing for their universal applicability and lack of dependence on the content's origin or embedded signals, they tend to have high rates of false positives and negatives, particularly as generative models improve and content undergoes minor edits or paraphrasing, limiting their use mainly to low-stakes environments (Sadasivan et al., 2025; Weber-Wulff et al., 2023). Retrieval-based approaches work by checking suspect content against a database of previously recorded AI outputs. These techniques are most effective when the reference databases are comprehensive and current, yet they face significant privacy and scalability hurdles, especially due to the need for extensive data storage and international governance (Krishna et al., 2023). The Information Technology Industry Council (ITI) also stress the importance of human authentication (a vital fail-safe despite being resource-intensive and subject to bias) while highlighting alternative approaches that leverage blockchain-based tracking and behavioural analysis (Council, 2024b).

Visible signals—while essential for promoting transparency and raising public awareness about AI-generated content (Burrus, Curtis, and Herman, 2024; Wittenberg, Epstein, Berinsky, and Rand, 2024)—are technically better described as explicit content labels or marks rather than watermarks in the strict sense, as now required in new regulations from multiple jurisdictions such as EU, US (some states) and China (Chomanski and Lauwaert, 2025; Fisher, 2024). Notwithstanding the well-versed debates on the efficacy, costs, and unintended consequences of labelling (Gamage, Sewwandi, Zhang, and Bandara, 2025; Jung, Hua, Bao, and Sundar, 2025; Scharowski, Benk, Küehne, Wettstein, and Brühlmann, 2023), these visible labels are similarly fragile. They can be quickly cropped, painted over, or otherwise erased using basic editing tools—often with minimal degradation to the underlying content. Once a labelled item is downloaded, reposted, or digitally repurposed, these explicit marks are frequently omitted, especially as files are converted between formats or screenshot mechanisms are employed to generate "clean" versions with no identifying tags. The distributed and often fragmented nature of the content sharing ecosystem further compounds this fragility: even if explicit labels exist at the point of first publication, their persistence cannot be assured as content migrates across platforms, intermediaries, and devices. In practice, any actor in the distribution chain can intentionally or inadvertently strip labels, rendering visible signals ephemeral.

Therefore, achieving robust provenance and authenticity for digital content is beset by persistent technical, organisational, and legal challenges. All methods—whether provenance tracking, watermarking, or human oversight—face trade-offs, and no solution is fool-proof; adversarial actors continually find ways to circumvent detection, making universal, perfect authentication unattainable. Hence, adopting a context-sensitive strategy that blends different techniques may represent the most realistic path forward. Critically, content authentication remains an adversarial, rapidly evolving field, demanding not just technical progress but also cross-platform cooperation, adaptive regulation, and shared accountability. As regulatory initiatives in China will reveal, sustainable solutions will require coordinated, ongoing adaptation rather than any single, definitive policy or technology.

3 China's AI Content Labelling Mandate: A Brief Overview

This section presents a succinct, high-level synopsis of China's Measure for Labeling of AI-Generated Synthetic Content (MLAIGSC), promulgated in March 2025 and in force since September 2025(**?**). China's hesitance to promulgate a comprehensive and omnibus legislative regime governing artificial intelligence is increasingly evident, as evidenced by the conspicuous absence of AI-specific regulation from its most recent legislative blueprints.[69] In lieu of sweeping statutory frameworks, China has opted

[69]The 2025 Legislative Work Plan of the Standing Committee of the National People's Congress (NPC) sets out comprehensive goals and priorities for the final year of China's 14th Five-Year Plan. In 2025, the work plan includes completed and ongoing review of legislation in areas such as economics, society, environment, and security (e.g., Infectious Disease Prevention Law, Financial Stability Law), as well as first reviews of 23 new laws covering topics like the Ecological Environment Code, Financial Law, and Social Assistance Law. Preparatory reviews are anticipated for major new domains—including State-Owned Assets Law, Excise Tax Law, and legislation on the healthy development of artificial intelligence and emerging technologies. The policy framework emphasizes Party leadership, legislative quality, people's participation, and alignment with national strategies. Other features include strengthened constitutional compliance review, international communication of legal texts, and optimization of the socialist legal system—especially targeting digital, ecological, foreign-related, and technological innovation legislation in 2025. See the full 2025 NPC legislative plan: http://www.npc.gov.cn/npc/c2/c30834/202505/P020250513550316685290.pdf (accessed 2025-07-31).The 2025 Legislative Work Plan of the State Council of China outlines the main legislative goals and priorities for the year. It focuses on advancing high-quality development, implementing the rule of law, and supporting initiatives in economic reform, social governance, public health, technological innovation, and environmental protection. The plan includes drafting and amending key laws and regulations related to digital economy, social security, ecological conservation, and risk management. Emphasis is placed on aligning legislation with strategic national objectives, enhancing public participation in the legislative process, and optimizing the legal

for a stratified approach—deploying a series of subordinate, mandate-oriented regulatory instruments that nonetheless wield substantial regulatory force. Enacted through the concerted action of a consortium of regulatory heavyweights—including the Cyberspace Administration of China, the Ministry of Industry and Information Technology (MIIT), the Ministry of Public Security, and the National Radio and Television Administration—this mandate signals an unequivocal intent to blanket the entire media and digital ecosystem, embedding compliance across diverse platforms and contexts. In recent years, China has developed a consistent pattern of imposing technology-specific regulations—progressing from algorithmic recommendation systems, through synthetic media controls, to the contemporary regime governing generative AI. The latest mandate on AI content labelling not only amplifies the horizontal sweep of regulatory oversight but also telegraphs a calculated bid for leadership in the global contest to frame, and perhaps dictate, the norms and protocols of next-generation AI governance. Far from deriving solely from indigenous ingenuity, China's evolving governance template is, at least in part, derivative—inflected by transnational currents, with discernible conceptual borrowings from ongoing legislative debates within the United States, even as its legislative machinery outpaces the more deliberative tempos characteristic of Western counterparts (Li and Chen, 2024). China's pragmatism manifests as a deliberate strategy, whereby emergent sectoral rules serve not only immediate domestic imperatives but also China's overt aspiration to consolidate normative influence over global AI standard-setting. This ambition has, in recent years, translated into a concerted effort to assert China's prominence well beyond academic discourse and into the international architecture of AI governance—most conspicuously crystallised in the trajectory of its AI Action Plan, which stands in notable juxtaposition to the contemporaneous, albeit more fragmentary, policy overtures of the Trump Administration in the United States(Guardian, 2025).

To start with, China's AI content labelling mandate operates a dual system that requires both explicit (visible) and implicit (metadata-based) labels for AI-generated content[70]; however, the regulatory framework currently provides no substantive provisions for tamper resistance or verifiable evidentiary mechanisms at the quasi-legislative level, with technical specifications relegated to standard-making processes that accompany the mandate.[71]

China's 2025 AI content labelling mandate, in contrast to the earlier 2023 "Deep Synthesis" regulation[72], explicitly prioritises metadata-based provenance tracking as an essential mechanism for authentication and traceability of AI-generated content, moving beyond the previous implicit framing.[73] The new regulations require all qualifying AI-generated content to contain both visible labels and embedded metadata, with the latter detailing creator and content attributes directly within the file's structure, thereby strengthening technical transparency and attribution throughout the content's lifecycle.

Digital watermarking is formally encouraged as a recognised authentication tool—but the 2025 rules remain conspicuously silent on the relationship between watermarking and metadata provenance, such as whether service providers can meet compliance using only watermarking or whether metadata-based tracking is strictly mandatory.[74] No regulatory guidance is offered on possible interoperability or substitution between these two approaches, leaving it uncertain if a provider may rely entirely on watermarking-based authentication without maintaining robust metadata provenance tracking within files.

---

framework for modernization. For details, see the full work plan: https://www.gov.cn/zhengce/content/202505/content_7023697.htm (accessed 2025-07-31).

[70] Art. 3, Measures for Labeling of AI-Generated Synthetic Content, defines two mandatory labeling types: explicit labeling—marks added to AI-generated or synthetic content in a manner easily perceptible to users (e.g., via text, audio, or graphics); and implicit labeling—technical markers embedded in the file data, not easily perceived by users. See: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm, translation at China Law Translate: https://www.chinalawtranslate.com/en/ai-labeling/ (accessed 2025-07-31).

[71] China has developed a comprehensive and mandatory national standard for AI-generated content labeling (GB45438-2025), which is set to take effect alongside a legal mandate. This standard is supplemented by technical guidelines such as TC260-PG-20233A and TC260-PG-20252A, providing detailed specifications and implementation pathways. As the effective date nears, additional guidelines continue to be published, focusing on the technical realization of invisible labeling in metadata across text, image, audio, and video content, as well as on security and technical inspection procedures. In contrast, efforts to create a national digital watermarking technical standard remain provisional; current drafts are conceptual and under public consultation, not yet reaching the operational detail or enforceability of the content labeling standards.

[72] The "Provisions on the Administration of Deep Synthesis Internet Information Services," adopted on November 25, 2022, establish requirements for labeling synthetic media, managing technical standards, and clarifying legal obligations for internet information services that use deep synthesis technologies in China. See the official text: https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm. For an English translation, see China Law Translate: https://www.chinalawtranslate.com/en/deep-synthesis/ (accessed 2025-07-31).

[73] Art 5 MLAIGSC

[74] Art. 5 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) encourages service providers to add forms of implicit labeling, such as digital watermarks, within generated synthetic content.

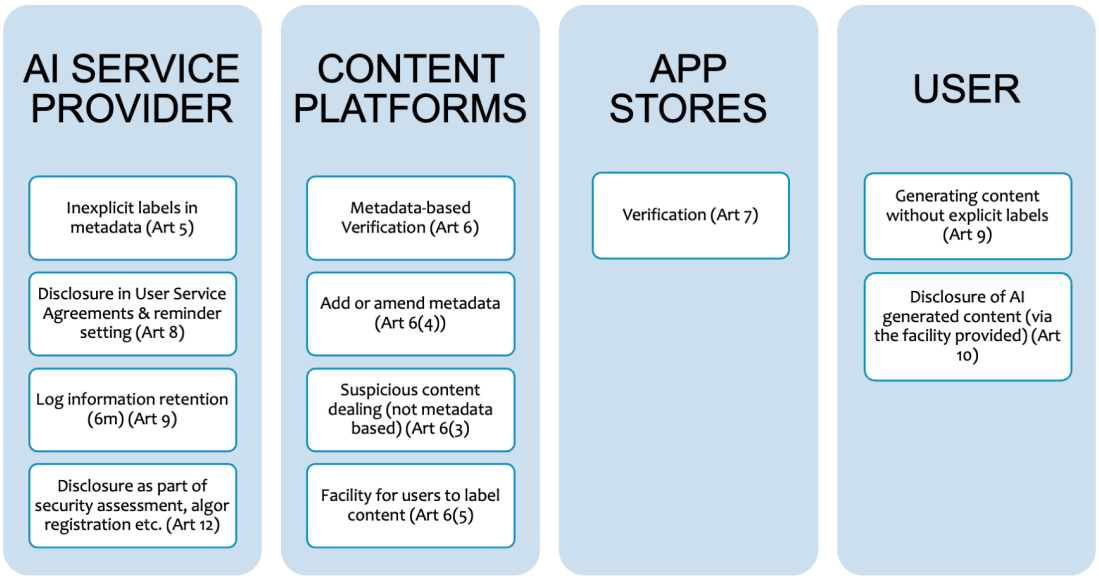

AI SERVICE PROVIDER
Inexplicit labels in metadata (Art 5)
Disclosure in User Service Agreements & reminder setting (Art 8)
Log information retention (6m) (Art 9)
Disclosure as part of security assessment, algor registration etc. (Art 12)
CONTENT PLATFORMS
Metadata-based Verification (Art 6)
Add or amend metadata (Art 6(4))
Suspicious content dealing (not metadata based) (Art 6(3)
Facility for users to label content (Art 6(5)
APP STORES
Verification (Art 7)
USER
Generating content without explicit labels (Art 9)
Disclosure of AI generated content (via the facility provided) (Art 10)

Fig. C1: Responsibility delineation across the content generation and dissemination chain under China's content labelling mandate

Most notably, China's AI content labelling mandate establishes a multi-layered system of attribution, assigning escalating obligations to all actors within the AI content pipeline—AI model developers, deployers/service providers, platforms, app stores, and end users—each interlocked to ensure continuous provenance and compliance.

AI service providers, including both developers and deployers, must embed both explicit (visible) and implicit (metadata) labels in all AI-generated content. The mandate bakes these requirements into their products "by design," disclose labelling practices in user agreements, and maintain generation logs for regulatory traceability.[75] Platforms, serving as enforcement intermediaries, are mandated to detect, classify, and supplement labels for any AI-generated content they host or distribute, even where upstream attribution may be missing or incomplete, and must embed their own metadata to reinforce or restore provenance throughout the content lifecycle.[76] App stores and distribution platforms are required to vet generative AI tools for labelling compliance as a condition of approval, screening for functional content labelling before allowing distribution to users.[77] End users must not remove, tamper with, or obscure any explicit or implicit content labels, and must actively declare whether content uploaded to platforms is AI-generated, utilising platform-provided labelling functions; however, when users request AI-generated content without explicit (visible) labels, service providers may accommodate this only after clearly assigning the user explicit labelling obligations and usage responsibilities through user agreements, and must, in such cases, retain relevant user and transaction logs for at least six months

[75] Arts. 4–5, 8 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) require service providers to add explicit labels to generated synthetic content if their service falls under Article 17(1) of the Provisions on the Administration of Deep Synthesis Internet Information Services, and to embed implicit identifiers within content metadata as prescribed by Article 16 of those Provisions. Providers must also clearly describe the labeling methods and specifications in user agreements, ensuring users are properly informed of the requirements. Notably, China's regulations use a broad definition of "service provider," making no distinction between actor roles in the AI supply chain—unlike the EU AI Act, which differentiates between "providers" and "deployers." As a result, a wide range of actors in China are subject to comprehensive transparency, safety, and accountability requirements regarding AI-generated content labeling.

[76] Art. 6 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) requires online content dissemination providers to regulate the spread of AI-generated synthetic content by: (1) checking file metadata for implicit markers and, if labeled as synthetic, adding prominent notices to alert the public; (2) if no marker is found but the user self-declares the content as synthetic, similarly adding prominent notices; (3) if neither marker nor user declaration exists, but explicit indicators or other signs of generated content are detected, treating the content as suspected synthetic and visibly notifying the public; (4) providing tools so users can declare synthetic content at publication. In cases (1)–(3), dissemination providers must also add information on content attributes, platform name/code, and content ID into the file metadata.

[77] Art. 7 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) stipulates that application distribution platforms must, during app listing or approval, require app service providers to state whether their applications offer AI-generated synthetic content services. If such services are provided, the distribution platform is responsible for verifying relevant documentation related to synthetic content labeling submitted by the app provider.

to ensure traceability and enforce accountability in line with regulatory requirements.[78] The reflects a systemic "chain of custody" approach to digital provenance and regulatory compliance, operationalised across all stages from content creation to distribution and end use, thus indicating a new standard of due diligence and potentially raised the bar for best practices.

A defining feature of China's regulatory paradigm in AI content authentication is its unswerving – and eminently pragmatic – fidelity to metadata-based provenance. This stance is animated by a pointed institutional skepticism regarding the present reliability of avant-garde authentication technologies such as advanced digital watermarking. Consequently, the 2025 mandate vests metadata labelling with cardinal regulatory significance, privileging traceability and operational lucidity over the premature adoption of nascent, insufficiently vetted technical solutions. Yet, the regulatory framework eschews rigidity in favour of strategic openness; it expressly encourages research and experiment with watermarking, intimating a readiness to integrate such innovations as their robustness is empirically demonstrated and their congruence with overarching policy objectives assured. In effect, China's approach secures immediate regulatory oversight and accountability through the institutionalisation of tried-and-tested metadata methods, while simultaneously preserving institutional latitude to adaptively incorporate more sophisticated mechanisms as technological circumstances evolve.

4 Chains of Accountability with Chinese Characteristics: Promises and Pitfalls

At a pivotal moment when the perils of provenance fragility and the proliferation of untraceable synthetic content threaten the structural integrity of digital trust, China's regulatory experiment stands as both a crucible and a lodestar for global governance. By weaving together meticulous institutional oversight with a patchwork of technical and organisational safeguards, China signals both a wary pragmatism and a willingness to grapple head-on with the many-hands dilemma that vexes accountability in our era of generative AI. (Jing, 2017). Examining China's evolving response is not merely instructive but essential, as its efforts to bridge the gaping chasm between regulatory aspiration and technical feasibility illuminate the formidable challenges—and latent possibilities—that confront all nations charting a course through this uncharted regulatory terrain.

In essence, China's strategy for addressing the inherent weaknesses of metadata-based provenance tracking in the regulation of AI-generated content is distinguished not by a wholesale embrace of cryptographically secure traceability technologies, as epitomised by several US state-level statutes, but by the institution of an organisationally rigorous and multifaceted governance framework. Rather than assuming that technological solutions alone—such as digital watermarking or cryptographic signatures—can resolve the fragility of metadata, Chinese regulators compensate for these vulnerabilities through an elaborate regime of layered institutional and platform-level obligations. This organisational diligence is reinforced by regulatory penalties for non-compliance, persistent audits, and the ongoing development of national standards—all designed to preserve an unbroken chain of accountability even as content circulates across diverse platforms and jurisdictions. (Cobbe, Veale, and Singh, 2023b)

First, China imposes comprehensive and interdependent obligations on all actors in the AI content pipeline—including service providers, platforms, app stores, and users—requiring not only the embedding and maintenance of metadata labels but also continual verification and supplementation at each interface. Platforms and intermediaries are mandated to actively detect, restore, or even replace missing provenance information, using a combination of automated checks, manual audits, and platform-level interventions. Where technical fragility arises (such as the corruption or removal of metadata), regulatory pressure compels these organisations to implement systems for reinforcing and remedying gaps, ensuring that provenance is recoverable and traceability is sustained across the ecosystem. (Longpre, Mahari, Obeng-Marnu, Brannon, South, Gero, Pentland, and Kabbara, 2024a). Importantly, while this regime prioritises organisational diligence and platform capabilities as bulwarks against the fragility inherent in metadata systems, it also consciously reserves regulatory space for the maturation of novel technological solutions, such as digital watermarking, that may in time offer a more definitive fix to the limitations of metadata alone.

[78] Arts. 9–10 of the Measures for Labeling of AI-Generated Synthetic Content (MLAIGSC) require users who publish AI-generated synthetic content through online dissemination services to proactively declare and label such content using the provider's labeling tools. If a user requests the service provider to issue synthetic content without explicit labeling, the provider may do so only after clarifying the user's labeling duties and responsibilities via user agreement, and must retain related logs—including the identity of the recipient—for at least six months to comply with legal requirements.

Furthermore, this organisational scaffolding is reinforced by a robust enforcement regime: non-compliance with labelling or provenance obligations exposes actors to unspecified regulatory penalties, thereby incentivising diligence and institutional self-policing. Service providers must maintain user logs and supply evidence upon request, while persistent monitoring and compliance audits seek to close potential accountability vacuums.

China's regulatory architecture also seeks to anticipate the many hands problem—where the diffusion of responsibility between numerous, diverse actors risks causing accountability vacuums. In response, the regulatory and technical approach aspires to "stitch together" an accountability chain that is seamless, even as content migrates across platforms and borders. Recognising the lack of cross-platform coordination, Chinese authorities proactively push for national and industry standards that aim to harmonise provenance protocols and lay groundwork for broader accountability solutions in the future. Therefore, unlike supply chain scholarship that mainly diagnoses how attribution evaporates as technical artefacts, decisions, or operational risks traverses a nested web of actors, each shielded by contracts, technical architecture and institutional boundaries(Nissenbaum, 1996)(Cobbe et al., 2023a), the China's regime imposes mandatory, persistent labelling at every stage, backed by regulatory penalties for all involved, not just originators. This attempt at operationalising provenance and shared responsibility also starkly contrast with the more voluntary, sectoral, or upstream-focused approaches seen elsewhere, foregrounding both the radical ambitions and new challenges of engineering accountability into the entire lifecycle of AI-generate content.

Despite the organisational rigour and interlocking regulatory safeguards, China's provenance regime is by no means free from blind spots. The framework exhibits limited consideration for the threat posed by malicious manipulation—where actors deliberately strip, alter, or falsify metadata and labels, subverting the system's traceability and impairing detection, or where AI-generated content is disseminated without any labeling via offshore, unofficial, or encrypted channels(Wang, Yan, Zhang, and Zhang, 2021). These are precisely the scenarios where provenance is most critical, yet the regime's reliance on metadata and platform-based controls may leave it ill-equipped to address.

Another significant blind spot in China's regulatory framework for AI authentication is the challenge posed by cross-jurisdictional dissemination. When such content moves transnationally—passing from a Chinese platform into Western social networks, or vice versa—the liability structures and technical capacities underpinning provenance often break down, especially considering the rigidity of China's emergent standards. The fragmentation is further exacerbated by a lack of global standardisation and coordination(Cihon, 2019a): metadata structures and authenticity conventions used in one legal environment may be unrecognised or unsupported in another, while visible indicators (such as overlays or icons) may be ignored or stripped at the point of import. These asymmetries have produced a fragmented global ecosystem for AI content authenticity, where regulatory effectiveness is constrained by the lack of mutual recognition, technical interoperability, and cross-border cooperation. In practice, national mandates cannot, in isolation, address these transnational governance deficits—a multilateral approach is indispensable. While it falls beyond the purview of this paper to undertake a detailed comparison, the juxtaposition of the emerging standards on AI content labelling put forth by China's TC260 (of the National Cybersecurity Standardization Technical Committee, 2025) and those evolving in Western jurisdictions (Krog, 2025) amply illustrates both the breadth of regulatory ambition and the formidable challenges attendant to harmonising—or developing—an international standard capable of ensuring effective cross-border provenance and accountability for AI-generated content. Given the routine movement of digital content across borders, regulatory fragmentation constitutes a profound and persistent problem(Cihon, 2019b). As China's experiment reveals, effective global AI governance must prioritise the mutual recognition and interoperability of provenance and authenticity standards, facilitate robust cross-border enforcement mechanisms, and nurture ongoing multilateral dialogue and consensus-building. International developments in this area have lagged the rapid proliferation of national mandates: coordinated solutions for aligning metadata standards, legal definitions, or technical protocols for AI content authentication across borders are still in their infancy. Despite being an apparent global governance problem, there remain scant initiatives within global or sub-global forums that systematically confront the legal and technical interoperability imperative.

While lauded for its thoroughness, China's regime for AI content labelling may draw criticism for its rigid, prescriptive, one-size-fits-all approach to provenance. The framework and supporting standards mandate uniform technical methods—dual explicit and implicit labelling, prescribed content placements, and highly visible notices in all media types—leaving little to no room for alternative mechanisms or

adaptation to sector-specific needs. This rigidity is especially pronounced in the technical standards that underpin the regulatory rules, which are both detailed and operationally mandatory, shaping not just policy but the minutiae of compliant technical implementation. Further, China's approach to AI-generated content labelling is marked by a lack of contextual nuance and flexibility. It mandates universal and visible disclosure of artificially generated content, offering little scope for discretion based on content type, intent, or potential for confusion among audiences. This stands in stark contrast to Article 50 of the EU AI Act, which explicitly considers the context and purpose of AI-generated content. For instance, Art 50(4) of the Act permits exceptions or more proportionate disclosure for "evidently artistic, creative, satirical, fictional or analogous" works, ensuring that mandatory transparency is implemented "in an appropriate manner that does not hamper the display or enjoyment of the work". Unlike the EU's model—which actively weighs context, intent, and audience expectations before mandating disclosure and allows for tailored implementation—China's approach notably centres on the risk of "public confusion" as the core trigger for mandatory disclosure. Under Art 17 of China's AI-generated content labelling mandate, platforms and providers must clearly label any synthetic content that could potentially "cause the public to confuse or misidentify its nature". Thus, it hinges upon a more rigid and categorical assessment of when content might mislead or confuse the public, with limited space for interpretive discretion or situational nuance. By contrast, the EU's approach leans towards a more nuanced, context-dependent assessment, enabling regulatory responses to be tailored to specific risks, audiences and environments. Further, the cumulative effect of ubiquitous labelling, provenance tracking, and log retention could produce chilling effects among users. If creators – particularly those working with AI tools in sensitive or contested cultural domains – fear misclassification, surveillance and regulatory sanction, they may begin to self-censor or withdraw content entirely. This is particularly true for hybrid content, where the line between AI-assisted and human-authored expression is internally negotiated and deeply contextual(Department, 2025). Pervasive governance carries the risk of flattening creative nuance and transforming AI-labelled content into a presumptively distrusted or stigmatised category, and regulators must balance the public interest in traceability with intellectual expression, especially in transmedia and user-generated environments(Treiger, 1989).

Lastly, there are other enforcement challenges rooted in the sheer scale, speed, and technical plurality of the AI content generation and dissemination pipelines. Ensuring consistent compliance across millions of user actions and a sprawling landscape of platforms necessitates not only detection infrastructure but human resources for auditing, nuanced legal interpretation, and unprecedented coordination across stakeholders. Although China has previously exhibited the capacity to orchestrate broad-scale regulatory interventions, the integrative complexity of AI-generated content—wherein metadata or labels can be trivially obfuscated, stripped, or manipulated—renders perfect enforcement out of reach. Compliance is thus likely to be selective, skewed toward the most visible actors: major tech platforms and prominent providers will be subject to close scrutiny and rigorous oversight, while content disseminated through informal, encrypted, decentralised, or cross-border networks will habitually escape detection. Further, social media and content sharing platforms—core enforcement nodes within this system—are tasked with proactively detecting and labelling AI-generated material. Yet, their detection capabilities are constantly tested by rapid technological advances, as well as by increasingly sophisticated evasion strategies employed by bad actors. These technical obstacles make enforcement a moving target, demanding ongoing research and development and the continual refinement of classifiers. Moreover, ramping up platform surveillance and detection capabilities can raise further issues regarding privacy, freedom of expression, and fairness. (Srinivasan, 2024). The resultant legal and social questions—of how detection systems operate, to what extent personal data is processed, and who bears responsibility for errors or overreach—remain unresolved. Moreover, enforcement is destined to be patchy, shaped by technical arms races and the persistent tension between regulatory ambition and the practical limits of control in a dynamic, borderless digital environment.

5 Conclusion

China's evolving legal architecture, despite the imprint of selective borrowing from both European and American approaches, represents a distinctive and under-examined intervention whose implications for global AI governance are profound but not fully mapped. While China's law is tailored to its domestic information ecosystem and administrative priorities, it catalyses a test case for global AI governance. Its ambition – operationalising provenance and collective responsibility at scale – exposes the tensions

between technological innovation, regulatory urgency, and compliance enforceability. As other jurisdictions weigh similar requirements, the Chinese approach offers both a template and a caution: technical mandates along will not resolve the epistemic opacity and unpredictable agency that defines digital media ecosystem, and a truly robust accountability architecture remains inherently unfinished – persistently challenged by both the arms race of forgery and detection, and the institutional politics of transnational content flow. In other words, China's AI content labelling mandate is not simply a regulatory milestone in the history of AI regulation, but a crucible wherein the limits and promise of engineered accountability are being tested – at the very forefront and on a global stage.

China's AI content labelling mandate neither sets a ready-made blueprint for other countries nor proposes a universally desirable path. Yet, its pragmatic, organisationally intensive focus—especially on downstream distribution—poses reflections Western governments must not ignore. While liberal democracies often trust technological advancement and authentication tools, they tend to leave a vacuum around regulatory rules at the very moment those technologies mature. Moreover, the industry's accelerated embrace of open-source models since DeepSeek redefined the competitive landscape further complicates the efforts to authentic and trace the origins of AI-generated content as it is nearly impossible to enforce standardised authentication requirements at the model or developer level(Danen, 2025). China's focus that heeds downstream dissemination thus emerges as a pragmatic response to these challenges.

This calculated move by China to foreground AI authentication signals not just a leap into the AI governance arena, parallel to the intensifying AI arms race, but functions as an useful reminder: The challenge is no longer just about inventing new safeguards but about ensuring that rules adapt alongside advances. As the legal and technological landscapes continue to evolve, the world's response to China's AI content governance experiment may define the next chapter in the global pursuit of attribution, accountability, and trust.

On the whole, China's approach is unlikely to gain traction in societies that prioritise individual liberties and pluralism. Still, it brings to light important questions about how global governance might address the seamless spread of AI-generated content across platforms and borders. Rather than aligning perfectly with current debates on AI supply chain accountability, the Chinese strategy partly diverges, engaging more directly with the control and distribution of content on social platforms and challenging us to consider how law can evolve in the face of constant technological change.

6 Acknowledgment

The author would like to express our sincere gratitude to the participants of the Paris Conference for AI and Digital Ethics, as well as the attendees of the SSPS Research Seminar at the University of Lincoln, for their insightful feedback and valuable contributions to the development of this paper.

7 Declaration

The author declares that there is no conflict of interest with respect to the publication of this article.

References

Adami M (2024) How ai-generated disinformation might impact this year's elections and how journalists should report on it. https://reutersinstitute.politics.ox.ac.uk/news/how-ai-generated-disinformation-might-impact-years-elections-and-how-journalists-should-report, accessed: 2025-07-31

Andrews C (2025) Virginia governor vetoes ai bill. https://iapp.org/news/a/virginia-governor-vetoes-ai-bill, published March 25, 2025. Accessed: 2025-07-31

Bereskin C (2023) Parliamentary handbook on disinformation, ai and synthetic media. https://www.cpahq.org/media/sphl0rft/handbook-on-disinformation-ai-and-synthetic-media.pdf, accessed: 2025-07-31

Burrus O, Curtis A, Herman L (2024) Unmasking ai: Informing authenticity decisions by labeling ai-generated content. Interactions 31(4):38–42. 10.1145/3665321, accessed: 2025-07-31

California (2024) Sb-942 california ai transparency act (2023–2024). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942, approved by Governor September 19, 2024; Effective January 1, 2026; Accessed: 2025-07-31

Cheney J, Chong S, Foster N, et al (2009) Provenance: a future history. In: OOPSLA '09: Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications, pp 957–964, 10.1145/1639950.1640064, accessed: 2025-07-31

Chomanski B, Lauwaert L (2025) Automated propaganda: Labeling ai-generated political content should not be required by law. Journal of Applied Philosophy 42(3):994–1015. 10.1111/japp.70002, first published: 24 February 2025. Accessed: 2025-07-31

Cihon P (2019a) Standards for ai governance: International standards to enable global coordination in ai research & development. Tech. rep., Future of Humanity Institute, University of Oxford, accessed: 2025-07-31

Cihon P (2019b) Standards for ai governance: international standards to enable global coordination in ai research & development. Future of Humanity Institute University of Oxford 40(3):340–342

Cobbe J, Veale M, Singh J (2023a) Understanding accountability in algorithmic supply chains. In: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, Chicago, IL, USA, pp 1186–1197, 10.1145/3593013.3594073, accessed: 2025-07-31

Cobbe J, Veale M, Singh J (2023b) Understanding accountability in algorithmic supply chains. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp 1186–1197

Commission UFC (2024) Fcc proposes disclosure rules for the use of ai in political ads. https://www.fcc.gov/document/fcc-proposes-disclosure-rules-use-ai-political-ads, notice of Proposed Rulemaking, FCC 24-74, Docket No. 24-211; Accessed: 2025-07-31

Council ITI (2024a) Authenticating ai-generated content: Exploring risks, techniques & policy recommendations. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf, accessed: 2025-07-31

Council ITI (2024b) Authenticating ai-generated content: Exploring risks, techniques & policy recommendations. https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf, accessed: 2025-07-31

Danen V (2025) The power of open collaboration: How open source is shaping the future of ai. https://www.forbes.com/councils/forbestechcouncil/2025/01/03/the-power-of-open-collaboration-how-open-source-is-shaping-the-future-of-ai/, published January 3, 2025. Accessed: 2025-07-31

Deelman E, Berriman B, Chervenak A, et al (2009) Metadata and provenance management. In: Shoshani A, Rotem D (eds) Scientific Data Management: Challenges, Technology, and Deployment. Chapman and Hall/CRC, accessed: 2025-07-31

Department HL (2025) Artificial intelligence and the creative double bind. Harvard Law Review 138(6):1585–1704. Chapter II.

EU (2024) Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), art 50. http://data.europa.eu/eli/reg/2024/1689/oj, official Journal of the European Union L, 2024/1689, 12.7.2024; Accessed: 2025-07-31

Feher K (2025) Generative AI, Media, and Society. Routledge, accessed: 2025-07-31

Fisher SA (2024) Something ai should tell you – the case for labelling synthetic content. Journal of Applied Philosophy 42(1):272–286. 10.1111/japp.12758, first published: 18 August 2024. Accessed: 2025-07-31

Gamage D, Sewwandi D, Zhang M, et al (2025) Labeling synthetic content: User perceptions of label designs for ai-generated content on social media. In: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp 1–29, 10.1145/3706598.3713171, accessed: 2025-07-31

Guardian (2025) China calls for global ai cooperation days after trump administration unveils low-regulation strategy. https://www.theguardian.com/technology/2025/jul/26/china-calls-for-global-ai-cooperation-days-after-trump-administration-unveils-low-regulation-strategy, published: 26 July 2025. Accessed: 2025-07-31

Hart TR, de Vries D (2017) Metadata provenance and vulnerability. Information Technology and Libraries 36(4):24–33. 10.6017/ital.v36i4.10146, accessed: 2025-07-31

House UW (2025) Winning the race: America's ai action plan. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf, july 2025. Accessed: 2025-07-31

Irwin K (2024) Openai adds labels to ai images, but says metadata 'easily' removed. https://uk.pcmag.com/ai/150789/openai-adds-labels-to-ai-images-but-says-metadata-can-be-easily-removed, published February 2, 2024. Accessed: 2025-07-31

Jing Y (2017) The Transformation of Chinese Governance: Pragmatism and Incremental Adaption. Governance: An International Journal of Policy, Administration, and Institutions 30(1):37–43. 10.1111/gove.12231, 30th Anniversary Essay; First published 20 July 2016

Jung Y, Hua P, Bao JA, et al (2025) Ai-generated or ai-modified? user reactions to labeling ai use in social media posts. In: CHI EA '25: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp 1–7, 10.1145/3706599.3720264, accessed: 2025-07-31

Kapoor S, Narayanan A (2024) We looked at 78 election deepfakes. political misinformation is not an ai problem. https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem, accessed: 2025-07-31

Krishna K, Song Y, Karpinska M, et al (2023) Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. Neural Information Processing Systems, pp 27469–27500, published: 10 December 2023. Accessed: 2025-07-31

Krog GP (2025) CEN-CENELEC JTC21 AI Standards: Complete Detailed Overview. urlhttps://jtc21.eu/wp-content/uploads/2025/06/CEN-CENELEC-JTC21-AI-Standards-Complete-Detailed-Overview.pdf, Comprehensive overview of the European AI standardization initiative and its alignment with the EU AI Act.

Li W, Chen J (2024) From brussels effect to gravity assists: Understanding the evolution of the gdpr-inspired personal information protection law in china. Computer Law Security Review 54:105994. https://doi.org/10.1016/j.clsr.2024.105994

Longpre S, Mahari R, Obeng-Marnu N, et al (2024a) Position: data authenticity, consent, & provenance for ai are all broken: what will it take to fix them? In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org, ICML'24

Longpre S, Mahari R, Obeng-Marnu N, et al (2024b) Data authenticity, consent, and provenance for ai are all broken: What will it take to fix them? MIT Generative AI Initiative (MIT-GenAI)

10.21428/e4baedd9.a650f77d, published March 27, 2024. Accessed: 2025-07-31

Mühlhoff R (2025) Chapter 11: Collective responsibility in the ethics of ai. In: The Ethics of AI: Power, Critique, Responsibility. Bristol University Press, 10.51952/9781529249262, accessed: 2025-07-31

of the National Cybersecurity Standardization Technical Committee S (2025) Notice on the issuance of six cybersecurity standards practice guidelines, including the "method for identifying ai-generated synthetic content — implicit metadata identification for text files". https://www.tc260.org.cn/front/postDetail.html?id=20250828165129, document No. [2025] 118, marked "Wangan Mizi"

Nissenbaum H (1996) Accountability in a computerized society. Science and Engineering Ethics 2:25–42. 10.1007/BF02639315

N.Y. (2024) Senate bill s7592a: Requires disclosure of the use of artificial intelligence in political communications. https://www.nysenate.gov/legislation/bills/2023/S7592/amendment/A, 2023–2024 Legislative Session, Sponsor: Senator Jake Ashby; Accessed: 2025-07-31

Pan B, Stakhanova N, Ray S (2023) Data provenance in security and privacy. ACM Computing Surveys 55(14s):1–35. 10.1145/3593294, accessed: 2025-07-31

Ren K, Yang Z, Lu L, et al (2024) Sok: On the role and future of aigc watermarking in the era of gen-ai. https://arxiv.org/abs/2411.11478, accessed: 2025-07-31, arXiv:2411.11478

Sadasivan VS, Kumar A, Balasubramanian S, et al (2025) Can ai-generated text be reliably detected? https://arxiv.org/abs/2303.11156, 10.48550/arXiv.2303.11156, accessed: 2025-07-31, arXiv:2303.11156

Scharowski N, Benk M, Kühne SJ, et al (2023) Certification labels for trustworthy ai: Insights from an empirical mixed-method study. In: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, pp 248–260, 10.1145/3593013.3593994, accessed: 2025-07-31

Schick N (2020) Deepfakes: The Coming Infocalypse. Twelve, New York, accessed: 2025-07-31

Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. ACM SIGMOD Record 34(3):31–36. 10.1145/1084805.1084812, accessed: 2025-07-31

Srinivasan S (2024) Detecting ai fingerprints: A guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, published January 4, 2024. Accessed: 2025-07-31

Taylor I (2024) Collective responsibility and artificial intelligence. Philosophy & Technology 37. 10.1007/s13347-024-00718-y, accessed: 2025-07-31

Treiger LK (1989) Protecting satire against libel claims: a new reading of the first amendment's opinion privilege. The Yale Law Journal 98(6):1215–1234

US A (2023a) Fact sheet: Biden-harris administration secures voluntary commitments from eight additional artificial intelligence companies to manage the risks posed by ai. https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/, published September 12, 2023. Accessed: 2025-07-31

US C (2023b) H.r.5586 - deepfakes accountability act, 118th congress (2023–2024). https://www.congress.gov/bill/118th-congress/house-bill/5586/text, introduced in the House of Representatives September 20, 2023; Accessed: 2025-07-31

US S (2025) Assembly bill a6540a: Stop deepfakes act (2025–2026 legislative session). https://www.nysenate.gov/legislation/bills/2025/A6540/amendment/A, requires synthetic content creation system providers to include provenance data on generated or modified synthetic content; Accessed: 2025-07-31

Wang X, Yan Z, Zhang R, et al (2021) Attacks and defenses in user authentication systems: A survey. Journal of Network and Computer Applications 188:103080

Washington (2025) Hb 1170 - 2025-26: Informing users when content is developed or modified by artificial intelligence. https://app.leg.wa.gov/billsummary?BillNumber=1170&Initiative=false&Year=2025, requires providers of generative AI systems to include detection tools and enable disclosure options for AI-generated or modified content; referred to House Rules Committee as of January 31, 2025; Accessed: 2025-07-31

Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al (2023) Testing of detection tools for ai-generated text. International Journal for Educational Integrity 19. 10.1007/s40979-023-00146-z, published: 25 December 2023. Accessed: 2025-07-31

Werder K, Ramesh B, Zhang RS (2022) Establishing data provenance for responsible artificial intelligence systems. ACM Transactions on Management Information Systems 13(2):1–23. 10.1145/3503488, published: 10 March 2022. Accessed: 2025-07-31

Wittenberg C, Epstein Z, Berinsky AJ, et al (2024) Labeling ai-generated content: Promises, perils, and future directions. An MIT Exploration of Generative AI: From Novel Chemicals to Opera 10.21428/e4baedd9.0319e3a6, published March 27, 2024. Accessed: 2025-07-31

Xu R, Hu M, Lei D, et al (2024) Invismark: Invisible and robust watermarking for ai-generated image provenance. https://arxiv.org/abs/2411.07795, accessed: 2025-07-31, arXiv:2411.07795

Łabuz M, Nehring C (2024) On the way to deep fake democracy? deep fakes in election campaigns in 2023. European Political Science 23:454–473. 10.1057/s41304-024-00482-9

Trust By Decentralization? Where Blockchain Systems Meet Responsible AI Challenges

Aggelos Kiayias[1]

[1] University of Edinburgh and IOG.

Corresponding author. E-mail: aggelos.kiayias@ed.ac.uk

Abstract

The importance of developing "Responsible AI" has recently soared as one of the key considerations in information technology (IT). It is thus pertinent to ask what insights one may glean from other research areas within computer science that have been focusing on solving related problems or building tools and methods for increasing the trustworthiness of IT systems and can be repurposed within this new context. Blockchain technology can be seen to intersect with the objectives of responsible AI and can be used as a foundational infrastructure for responsible AI development. To highlight this direction, we briefly outline four use cases: decentralized identity for AI systems, blockchain-based governance, privacy-enhanced data mining, and decentralized machine learning. We conclude by discussing a "rights by design" approach for responsible AI development and with an overview of the papers in the session.

Keywords

Responsible AI, Blockchain, Governance, Privacy-Enhanced Data Mining, Decentralized Machine Learning.

1 Introduction

Lawrence Lessig's tenet that "Code is law" highlights how software and algorithms influence behavior in cyberspace, much like legal systems do the same thing in the physical world Lessig (1999). AI models, in this context, represent a significant challenge due to their (i) ability to influence people in a very direct manner, e.g., via interaction on the basis of text or avatar based communication (ii) potential for opaqueness, with large models being difficult to analyze or understand, (iii) unreliability, with AI models being vulnerable to biases, hallucinations and even misclassification attacks especially when reacting to maliciously crafted inputs, see e.g., Elsayed, Shankar, Cheung, Papernot, Kurakin, Goodfellow, and Sohl-Dickstein (2018). It follows that as we integrate AI into critical sectors, the consequences of AI failures escalate, making the need for AI safety an imperative.

To mitigate the risks regarding AI, a set of principles have been assembled that epitomize what a "Responsible AI" system might look like. These pillars can be seen as "north stars" in the effort towards developing trustworthy AI systems.

- **Fairness:** Ensuring that AI systems do not perpetuate or amplify existing societal biases, which may be inadvertently introduced during training.
- **Transparency:** The reasoning behind a particular output of an AI system is understandable and allows for meaningful explanation and scrutiny at a level where the end-user can understand.
- **Accountability:** The system is endowed with mechanisms that enable users to challenge, audit, and assign responsibility for failures or incorrect outcomes.
- **Governance:** Representative and inclusive decision making processes are in place that facilitate correcting, and suitably evolving the AI system.
- **Privacy:** Sensitive and personal data should be protected from unauthorized exposure or exploitation throughout the AI lifecycle.
- **Reliability:** The AI system should operate following the intent of the operator without deviating from reasonable and appropriate output or causing harm.

It is important to highlight that the above considerations permeate more broadly Information Technology (IT) systems. Blockchain technology is a recent development that suggests a paradigm shift from centralized IT to decentralized IT systems. This shift is fundamental to moving from "Rule by Code," where code is dictated by a single entity or organization, to a setting of "Rule of Code," where the rules specified in the software are themselves transparent, immutable, and collectively governed De Filippi, Mannan, and Reijers (2022). Blockchain technology, by distributing control, can provide the foundational infrastructure that may act as a substrate for future responsible AI. To illustrate this potential in the next section we will review four use cases that highlight how blockchain technology may support the objectives of responsible AI.

2 Use Cases at the Intersection of AI and Blockchain

2.1 The Right to Robot Identification

A critical question in the context of human to AI interaction is whether people have the right to know they are interacting with an AI and furthermore whether they are able to have the AI model's "fingerprint" recorded for future reference as well as authenticate the model's credentials. In a high-risk setting, recent regulation such as European Union's AI Act mandates that disclosure is required and impact assessments should be performed by providers, cf. Kusche (2024). A decentralized identity infrastructure (cf. Avellaneda, Bachmann, Barbir, Brenan, Dingle, Duffy, Maler, Reed, and Sporny (2019)), built on a public blockchain, can support this by:

- Creating a provable audit trail of AI interaction that may be referred to later, in case the human user wishes to challenge a specific outcome.
- Allowing AI entities and agents to acquire certified attributes and demonstrate such attributes reliably when interacting with users.
- Using Zero-Knowledge Proofs it is possible to verify fairness and accuracy of classification without revealing proprietary model details.

For more details on how the above features become relevant in providing accountability of AI from a legal perspective one can refer to Georgiou and Kiayias Georgiou and Kiayias (2020).

2.2 AI Governance

Governance of AI systems comes with an array of significant challenges, cf. Taeihagh (2021). Stakeholders, in the broadest possible sense, should have the opportunity to provide feedback and influence the system in a representative manner. This parallels blockchain-based governance that has recently embraced "digital constitutionalism" Mannan, De Filippi, and Reijers (2024) and facilitates encoding stakeholder roles and responsibilities, as well as provides methods for collective decision-making. This approach facilitates:

- **Transparency and Verifiability:** Governance decisions and their outcomes are recorded on an immutable ledger.
- **Privacy:** Stakeholder inputs, such as votes, can be kept private and coercion-resistant using cryptographic protocols.
- **Accountability and Sustainability:** Incentives are aligned so that they ensure the long-term health and responsible development of the system.

For further context on the above one can refer to Kiayias and Lazos Kiayias and Lazos (2022). Some examples drawn from the blockchain setting are the Cardano constitution and governance model[79] and the Polkadot governance framework.[80]

2.3 Privacy Enhanced Data Mining

AI models, especially those aiming at generative AI, typically require vast amounts of data for training. This raises privacy and copyright considerations that can pose significant barriers to model development. Moreover, there are significant volumes of data that are not typically publicly available and hence remain inaccessible for AI model training.

Using a blockchain as a registry for datasets and processing requests, it would be feasible to employ Secure Multiparty Computation (MPC) to enable data owners or controllers to engage in training without revealing private information. We note that training models in a privacy-enhanced fashion using cryptography is something that has already been studied at some length, e.g., see Bonawitz, Ivanov, Kreuter, Marcedone, McMahan, Patel, Ramage, Segal, and Seth (2017) and Bagdasaryan, Kairouz, Mellem, Gascón, Bonawitz, Estrin, and Gruteser (2022). Furthermore, it would be also possible to maintain a model secret shared amongst a set of servers and access it in a distributed manner without revealing the model itself to any individual participant; this would amount to a federated learning system Kairouz, McMahan, Avent, Bellet, Bennis, Bhagoji, Bonawitz, Charles, Cormode, Cummings, D'Oliveira, Eichner, El Rouayheb, Evans, Gardner, Garrett, Gascón, Ghazi, Gibbons, Gruteser, Harchaoui, He, He, Huo, Hutchinson, Hsu, Jaggi, Javidi, Joshi, Khodak, Konecný, Korolova, Koushanfar, Koyejo, Lepoint, Liu, Mittal, Mohri, Nock, Özgür, Pagh, Qi, Ramage, Raskar, Raykova, Song, Song, Stich, Sun, Suresh, Tramèr, Vepakomma, Wang, Xiong, Xu, Yang, Yu, Yu, and Zhao (2021) that is cryptographically secured. The role of the blockchain in this setting is one of access control: it can track data sets and processing requests so that it is ensured that the use of the model does not jeopardize the privacy of the underlying data sets.

2.4 Decentralized Machine Learning

Expanding on the last use case on the direction of collaborative model training, one can envision a setting where the whole AI lifecycle is realized as a resource based system following the framework outlined in Kiayias (2022). In particular, this means that nodes collaboratively train a global AI model on their local, heterogeneous datasets in a way that is incentivized and collectively managed. Using blockchain techniques a tokenized economy around the service emerges so that:

- Contributions for training and model improvement as well as computational power become tokenized and they are rewarded with a digital currency.

[79] See https://cardano.org/governance/.
[80] See https://polkadot.com/opengov/.

- Tokens are used for accessing the model; this facilitates an economic cycle that can incentivize participation and system maintenance indefinitely without a centralized authority.

It is worth noting that similar approaches to building IT systems with suitable "cryptoeconomics" have been developed in other cases outside the cryptocurrency spacce, see, e.g., Diaz, Halpin, and Kiayias (2022) for the case of mix-nets. While different from AI, this demonstrates how building incentives around digital tokens is feasible for different types of IT systems.

3 AI via "Rights by Design"

Blockchain technology and cryptography point to a more principled way to think about the design of AI (and IT in general) systems: follow a "rights by design" approach. This is inspired by the "security by design" principle which recognizes that cyber security cannot be an afterthought in system development.[81] In the same way, we can argue that basic rights cannot be left for the later stages of system development in AI. To follow this approach to system design and deployment, instead of starting with the technological objectives first, we ought to start with the fundamental rights that should be upheld by the system. For instance, such rights include data privacy, data portability, the right to erasure, the right to be informed, and freedoms of speech and association.

Viewing AI development from this lens, it is also possible to highlight the potential benefits brought forth by blockchain technology. Blockchain systems, by removing single points of failure for all relevant properties, can form a "neutral" substrate over which AI systems can be architected in a more responsible manner using basic rights as the guiding principle.

4 Session overview

In this session, there were three papers that were selected to appear and are included in these proceedings. We summarize them below.

- **From Black Box to Glass Box: Leveraging Blockchain to Audit AI Systems Through Multistakeholder Participation** by Roberto Almeida and Lucas Latini.

  This paper proposes a novel framework to address the "black box" problem in AI, where the opacity of AI systems creates a deficit of trust and accountability. The authors advocate for using blockchain technology to transform AI systems into transparent "glass boxes" open to examination. The core of the framework is a blockchain-based audit trail that stores immutable, cryptographically verifiable proofs of an AI's entire lifecycle. This includes records of data sources, model parameters, algorithmic updates, and inference events.

  Their proposal strengthens previous work by introducing three key principles: (i) Multistakeholder Governance: It enables diverse actors—such as regulators, academics, and civil society—to collectively validate an AI system's integrity, democratizing oversight and preventing control by a single entity. (ii) Selective Transparency: A layered access model balances public transparency with the need for confidentiality, granting different stakeholders access to different levels of information. (iii) Cryptographic Verifiability: By using hashes and digital signatures, the framework ensures that all audit records are tamper-proof and can be independently verified by third parties. Ultimately, the framework aims to create a more trustworthy and legally compliant AI ecosystem by providing a technical foundation for verifiable and socially accountable auditing.

- **Building Trust and Developing Value between Creative Rightsholders and GenAI through Tokenised Licensing and Content Authenticity Tools** by Chris Elsden, Caterina Moruzzi, Burkhard Schafer, Frances Liddell, Ella Tallyn, Evan Morgan, Billy Dixon, Kar Balan and John Collomosse.

  This paper addresses the significant trust deficit and regulatory gap between creative rightsholders and the developers of Generative AI (GenAI) systems that use their content. It proposes and critically evaluates a technical approach to build trust and facilitate fair value exchange. The core of the paper is an examination of the ORA (Ownership, Rights, Attribution) framework. This system is presented as a potential solution to give creators more control over their work. It combines two key technologies: (i) Content Authenticity (C2PA): This open standard provides verifiable creation provenance by embedding a manifest within a media file. This manifest securely records details like the original

[81] See e.g., the UK government's "secure by design" policy, https://www.security.gov.uk/policy-and-guidance/secure-by-design/policy/.

author, software used, and any subsequent edits or manipulations. (ii) Tokenized Licensing: Using a Distributed Ledger Technology (DLT), the framework creates programmable and transferable licenses as tokens. This allows for tracking ownership provenance (who owns the asset) and enables creators to set granular, machine-readable permissions for their work.

The authors argue against a one-size-fits-all solution, emphasizing that different creative industries have vastly different norms regarding collaboration, remixing, and intellectual property that any technical system must respect. Tokenization offers the potential for bespoke, automated licenses that go beyond a simple "opt-in" or "opt-out" system. Creators could specify conditions related to particular AI models, usage types (e.g., training vs. fine-tuning), or even set up automated royalty splits for collaborative works. The paper suggests that tdecentralized tools could support new forms of collective representation for creators, such as DAOs (Distributed Autonomous Organizations), which could manage licensing and compensation on behalf of a community. The authors also acknowledge significant limitations, including the technical complexity and usability barriers of DLTs, the need for robust identity systems to prevent false attribution, and the risk that bespoke licenses could become overly restrictive in derivative works. The article ends by presenting a forward-looking interdisciplinary research agenda with key questions about technology, governance, and market models needed to build a more sustainable and equitable creative ecosystem in the age of AI.

- **Bootstrapping Trust in Web3 from Publicly-Verifiable Internet Data**, by Yuan Lu, Qiang Tang.

  This paper introduces a novel *oracle protocol* to bridge the trust gap between the established, centralized Web2 systems and the decentralized Web3 ecosystem. It addresses a key limitation in previous state-of-the-art systems like DECO, which rely on a designated verifier to validate data from a TLS session. This reliance creates a central point of failure and prevents true public verifiability, as a malicious verifier can strategically abort the protocol, preventing proof generation.

  To solve this, the authors propose a new method called the "encrypt-TLS-keys-to-future" paradigm. The protocol works as follows: (i) Joint TLS Emulation: A prover (user) and a verifier (oracle service) use secure two-party computation (2PC) to jointly act as a single TLS client. Through this process, they each receive an additive share of the TLS session key, but neither party learns the full key. (ii) Time-Locked Encryption: The parties then use a trapdoor verifiable delay function (tVDF) to encrypt the full TLS session key. This encrypted key is only decryptable after a predetermined amount of time has passed. (iii) Public Proof Generation: The prover commits the encrypted TLS traffic to the blockchain. After the time delay expires, the prover can solve the VDF to reconstruct the full session key. They can then use zero-knowledge proofs (zk-SNARKs) to prove the authenticity of specific data from the TLS session without revealing anything else.

  A key advantage of this design is a novel form of fairness. It guarantees that the prover can generate a proof for any data that was already transmitted during the TLS session, even if the verifier becomes unresponsive or maliciously aborts the protocol. To demonstrate its utility, the authors show how this protocol enables the first decentralized fair exchange between Web2 assets (e.g., fiat currency) and Web3 assets (e.g., cryptocurrency) without needing a trusted intermediary.

I am grateful to Dr. Morshed Mannan, Dr. Thomas Zacharias and Prof. Burkhard Schafer for their reviewing all submitted works and contributing to the track's program selection.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Avellaneda O, Bachmann A, Barbir A, et al (2019) Decentralized identity: Where did it come from and where is it going? IEEE Communications Standards Magazine 3:10–13

Bagdasaryan E, Kairouz P, Mellem S, et al (2022) Towards sparse federated analytics: Location heatmaps under distributed differential privacy with secure aggregation. Proceedings on Privacy Enhancing Technologies 2022(4):162–182

Bonawitz K, Ivanov V, Kreuter B, et al (2017) Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. ACM, pp 1175–1191

De Filippi P, Mannan M, Reijers W (2022) From rule by code to the rule of code: A new framework for the rule of law in the digital age

Diaz C, Halpin H, Kiayias A (2022) Reward sharing for mixnets. Cryptoeconomic Systems 2(1). URL https://doi.org/10.21428/58320208.00acf460

Elsayed GF, Shankar S, Cheung B, et al (2018) Adversarial examples that fool both computer vision and time-limited humans. In: Advances in Neural Information Processing Systems (NeurIPS), pp 3914–3924, URL https://arxiv.org/pdf/1802.08195.pdf

Georgiou L, Kiayias A (2020) The right to robot identification: Using cryptographic hashes to manage ai black-boxes. URL https://ssrn.com/abstract=3776726

Kairouz P, McMahan HB, Avent B, et al (2021) Advances and open problems in federated learning. Foundations and Trends in Machine Learning 14(1–2):1–210

Kiayias A (2022) Decentralizing information technology: The advent of resource based systems. In: Algorithmic Game Theory - 15th International Symposium, SAGT 2022, Colchester, UK, Proceedings, Lecture Notes in Computer Science, vol 13584. Springer, pp 3–19

Kiayias A, Lazos P (2022) Sok: Blockchain governance. In: 4th Advances in Financial Technologies (AFT) 2022, pp 61–73

Kusche I (2024) Possible harms of artificial intelligence and the eu ai act: Fundamental rights and risk. Journal of Risk Research pp 1–14. URL https://doi.org/10.1080/13669877.2024.2350720

Lessig L (1999) Code and Other Laws of Cyberspace. Basic Books

Mannan M, De Filippi P, Reijers W (2024) The emergence of blockchain constitutionalism. In: De Gregorio G, Pollicino O, Valcke P (eds) The Oxford Handbook of Digital Constitutionalism. Oxford University Press, URL https://doi.org/10.1093/oxfordhb/9780198877820.013.46

Taeihagh A (2021) Governance of artificial intelligence. Policy and Society 40(2):137–157. URL https://doi.org/10.1080/14494035.2021.1928377

Building Trust and Developing Value between Creative Rightsholders and GenAI through Tokenised Licensing and Content Authenticity Tools

Chris Elsden[1], Frances Liddell[1], Caterina Moruzzi[1], Burkhard Schafer[2], Ella Tallyn[1], Evan Morgan[1], Billy Dixon[1], Kar Balan[3], John Collomosse[3]

[1] Institute for Design Informatics, University of Edinburgh, Edinburgh, United Kingdom.
[2] School of Law, University of Edinburgh, Edinburgh, United Kingdom.
[3] CVSSP, University of Surrey, Guildford, United Kingdom.

Corresponding authors. E-mails: chris.elsden@ed.ac.uk; fliddell@ed.ac.uk; cmoruzzi@ed.ac.uk; B.Schafer@ed.ac.uk; E.Tallyn@ed.ac.uk; E.Morgan@ed.ac.uk; B.Dixon@ed.ac.uk; k.balan@surrey.ac.uk; j.collomosse@surrey.ac.uk

Abstract

There is currently a deficit of trust between the developers of AI systems and creative rightsholders upon whose content many AI systems ultimately depend — and currently exploit. The rapid development of GenAI tools, in particular, leaves a regulatory gap, where traditional copyright and licensing models are ill-equipped to both protect rightsholders and support sustainable and ethical AI development.

In this paper, we examine the extent to which decentralised, tokenised licensing systems — specifically the ORA framework — in combination with provenance and content authenticity initiatives (e.g., C2PA), can address these needs. As a UK-based interdisciplinary research team across design, computer science, and law, we have engaged creative professionals and publics in workshops, demonstrations, and exhibitions related to the ORA framework and content provenance.

Reflecting on these together in this paper, we critically assess the role these systems could play in building creatives' trust in AI systems which use their work, and the further questions they raise. Finally, we present an interdisciplinary research agenda to stimulate further critical research on these technologies and their applications in the context of GenAI.

Keywords

Tokenised Licensing, Content Authenticity, GenAI, Creative Industries, Copyright, Design Research.

1 Introduction

There is currently a deep deficit of trust between the developers of AI systems, and various creative rightsholders upon whose content many AI systems ultimately depend and largely exploit (Authors Licensing and Collecting Society, 2024; Daly et al., 2025). Creatives currently have limited knowledge or control over how their work is being used as data to train AI models; few means to express consent, claim credit, or gain compensation related to AI-generated works and then may find themselves competing with machines which can be used to trivially reproduce a similitude of their work. For many creatives already working in a precarious sector of the economy (Brook, O'Brien, and Taylor, 2020), the rapid evolution of generative AI (GenAI) hence appears deeply threatening, requiring steadfast resistance to protect the futures of their creative practice.

On the other side, as a general-purpose technology, economies, governments, companies and individuals are deeply invested in a race to develop and deploy AI systems in ways that are hoped will lead to productivity and economic growth (Government, 2025). However, AI depends fundamentally on a rich supply of data (and computing power and energy). The largest foundation models require vast amounts of data for initial training; and then may be further *fine-tuned* using additional data, to produce certain output, and finally, specific prompts and data input are provided at the point of *inference* – when an output or work is ultimately produced (Brown, 2023).

The quality, labelling, originality, and contemporaneity of this data are crucial. Training a GenAI model on synthetic data can degrade the output (Shumailov, Shumaylov, Zhao, Papernot, Anderson, and Gal, 2024), and hence the constant (and necessary) improvement of AI models will depend on a ready and well-understood supply of new data, as well as confidence in the legality and sustainability of these data sources. Hence, while AI model developers appear to present a threat to creatives rightsholders, they also depend upon their past and future content production. This dilemma is illustrated around the world through ongoing legal battles over AI, copyright and data mining, and most recently in the UK, in vociferous debates and public consultation on how to facilitate AI companies access to data while protecting the UK's creative industries[82].

Beyond this adversarial perspective – of struggling creatives vs empire-building AI companies – it is important to consider the wider implications for the digital commons, and public, cultural infrastructures and ecosystems. Keller (Keller, 2025) argues powerfully that the primary goal of legislators should be to develop a sustainable public information infrastructure, *"…to ensure that these crucial societal functions—information production, maintenance, and accessibility—remain viable in an environment increasingly shaped by generative AI."* Where certain kinds of creative work or content is withheld, no longer produced, or unevenly distributed, AI models are likely to perpetuate further bias and harm (Huang and Siddarth, 2023). However, we should also recognise how many artists seek to use AI tools to as part of their practice, and that not all data in an AI model is equal. Data used in *fine-tuning*, or at the point of *inference* can have significantly more impact than data used in foundational training when tailoring a model to a specific domain or task.

Major rightsholders, such as media companies and publishers are exploring licensing deals with AI companies to allow access to their content and collections[83]. However, there is much valuable creative content which falls outside of traditional copyright, and – just as is the case in music streaming – the extent to which the benefits of such deals flow down to individual artists of creatives can be limited (Marshall, 2015). Considering the broad-based and individual nature of the 'creator economy', it is also evident how creatives of all kinds increasingly depend upon online platforms to share their work, connect with audiences, and possibly get paid (Duffy et al., 2019'). Many creators hence face a dilemma between making their work discoverable and enduring the terms and conditions of the online platforms they depend upon to do so (Nieborg and Poell, 2018).

In many countries, the interpretation and enforcement of copyright, as it relates to data mining for AI training, has become a central battleground between AI companies and creative rightsholders [84]. In the UK specifically, a recent consultation and public debates have centred on whether, and how, rightsholders may 'opt-out' or 'opt-in' to the use of their content to train AI models [see footnote p2]. While some larger rightsholders and publishers have made specific commercial deals to license their content or collections, the legal implications of AI training or reproducing aspects of copyrighted material are unfolding in the

[82] https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence

[83] https://petebrown.quarto.pub/pnp-ai-partnerships/

[84] Running list of lawsuits updated here: https://chatgptiseatingtheworld.com/2024/08/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/

courts in various jurisdictions. However, others have cautioned a 'Copyright Trap' (Craig, 2024) and argued that: **a)** copyright protections alone are not sufficient to produce a sustainable public information ecosystem (Keller, 2025); **b)** copyright may be of limited effectiveness (given how much data AI systems can already access which is under copyright); and **c)** there is a potential over-reach of copyright, in ways that may further entrench the power of a smaller number of larger rightsholders. Copyright will clearly remain relevant in some cases, in relation to some aspects of how GenAI is developed and applied, but it is unlikely to be a holistic solution to the current concerns.

Whatever the proposed regulatory solutions to these challenges, currently, many individual creators feel powerless, not just vis-a-vis AI developers, but also the large rights holders whom they have assigned their rights to (e.g. publishers, social media platforms). This lack of trust and feeling of disempowerment can have harmful consequences for society, from creators withdrawing entirely to forms of direct resistance and obstruction (e.g. Nightshade [85]) which could have wider unintended consequences and harm for an open and accessible internet.

One approach to restore trust is to enable creators to claim and prove their authorship and ownership of their work, then express and signal their intentions and preferences in a way AI companies are required to take notice of. If, as it is likely in the UK and EU, either an opt-in or opt out approach is chosen, this expression would also be necessary to enable appropriate attribution and compensation. There is a technical opportunity to develop tools and systems by which rightsholders, both large and small can express these intentions. Given the scale of data mining and processing, this information needs to be machine readable, trustworthy and support high volumes of transactions that can be executed somewhat automatically. In addition, from the perspective of AI developers and users, having reliable data about the content or media on which it is trained, and ingests, is increasingly important.

The paper now proceeds as follows. In Section 2 we briefly set out promising technical approaches to enable creators to prove authorship and express rights and intentions; focusing specifically on standards for provenance data, content authenticity and tokenised licensing. We then introduce our interdisciplinary and design-led approach to explore and interrogate what creators themselves, and the general public think about these approaches, and identify emergent issues and questions for further research with a view to building trust and value in an open information ecosystem.

2 Technical Developments and Opportunities

In context of the issues outlined in the introduction, the recently published ACCCT framework (Access, Consent, Control, Compensation, Transparency) presents one promising approach to reconcile the needs of AI developers and creative rightsholders in the era of generative AI (Bennett et al., 2025). Based on engagement from academics and industry stakeholders, the report recommends building upon emerging open standards for provenance and rights declaration and combining these with decentralised registries to attach durable metadata to media assets that carries granular, machine-readable signals for consent and attribution. The underpinning computer science for these layered approaches is beyond the scope of this paper and only a brief overview is provided below, however they are described comprehensively in prior work by co-authors Balan and Collomosse (Balan et al., 2023a; Balan, Gilbert, Black, Jenni, Parsons, and Collomosse, 2023b; Balan, Gilbert, and Collomosse, 2025; Collomosse and Parsons, 2024).

2.1 Content Authenticity and Provenance Data

There has been decades of research in computer science into content identification schemes to support content authenticity, including watermarking (which actively embeds metadata within digital media in ways that are difficult to remove) and fingerprinting (which passively derives identifiers from the media's inherent characteristics), approaches which can support robust downstream tracking through various transformations. While no single solution is perfectly secure, Collomosse and Parsons (Collomosse and Parsons, 2024) describe how these techniques can be combined to produce a durable method for attaching verifiable provenance facts to media assets, such as images, videos or audio. These facts – stored in tamper-evident metadata called a manifest – could include details about when media was created, the software and hardware used, authorship, and records of any manipulations of the media.

In practice, this field of research has contributed to the development of a widely adopted technical standard called 'Content Credentials', as part of the Coalition for Content Provenance and Authenticity

[85] https://nightshade.cs.uchicago.edu/whatis.html

(C2PA) [86], a cross-industry working group formed under the Linux Foundation and steered by Adobe, Amazon, BBC, Google, Intel, Meta, Microsoft, OpenAI, Publicis Groupe, Sony and Truepic, and is being developed as an ISO standard under ISO/IEC 21617-1:20. This standard can be used to identify changes to visual media as it circulates online, and is also widely used to label images produced by GenAI. These credentials do not directly represent claims about truth – they are a tool to judge media – however they do provide distinct provenance facts which can subsequently be used to discuss the authenticity, meaning or value of media (Moruzzi et al., 2025). The broad adoption of this standard indicates the growing significance of provenance data in a world of increasingly synthetic media, and the potential for media to be distrusted where verifiable provenance is not demonstrable.

2.2 Tokenised Licensing

Managing intellectual property (IP) using tokenisation and distributed ledger technologies (DLTs) has been an established use case in the creative industries. Early examples include the partnership between Christies and Artory from 2018, where DLTs were used to verify the data and provenance of the Barney A. Ebsworth collection. However, tokenisation alone is limited in its use as a registry system because ownership provenance and creation provenance are not linked to the non-fungible token (NFT) (Balan et al., 2023a), meaning claims can be falsely made about the rightful owner of the asset. Moreover, not all forms of tokenisation represent the copyright and IP of the asset. This is further complicated by the lack of standardisation over what a token represents when purchased from an NFT platform (Noh et al., 2023).

The recent growth in infrastructure and protocols leveraging DLTs for IP points to the significance of this technology in providing a tamper-resistant record and distribution of IP rights for digital assets (e.g. Ocean Protocol, Story Protocol). Tokenised licensing is part of this broader infrastructure of integrating IP with DLTs by using tokens to create transferable and programmable licensing that are separate from the digital asset.

2.3 The ORA Framework

In our research, we have focused on just one specific implementation of tokenised licensing – the ORA (Ownership, Rights, Attribution) framework. ORA combines the affordances of tokenisation with the C2PA (Balan et al., 2023a). By building directly upon C2PA, ORA provides a nuanced incorporates both ownership provenance (the exchange of media ownership) recorded on a DLT, and creation provenance that records the changes and edits made to a piece of media, including the attribution of the original creator and editors, via a C2PA manifest which is embedded into the digital media file. In ORA, standard NFT protocols (ERC-721) are used to create sets of tokens that separately represent the asset and licenses granted to it. Proposals for hierarchical NFT protocols (e.g. EIP-6150) might also be used for scalability. Crucially, when digital media are remixed from, or a composite with, other media, such as AI generated work or mashups, the sources and provenance of these media can be tracked and visualised to demonstrate the historical context of the work via a C2PA manifest. Therefore, the adoption of the C2PA standard is a distinctive feature of ORA in comparison to other protocols.

3 Methodological Approach

3.1 Studying ORA through an Interdisciplinary Design Research Programme

As authors, we represent an interdisciplinary team of researchers, across the fields of design, computer science and law collaborating through DECaDE [87] a large publicly-funded (UKRI/EPSRC) applied research programme to develop, explore and critically reflect on the decentralised digital economy. Computer scientists (Balan and Collomosse) have led development of several novel techniques and protocols (including Balan et al. (2025) and Balan et al. (2023b)), as well as the ORA framework (Balan et al., 2023a), and contributed to key national proposals and standards, such as the 'Time to ACCCT' report (Bennett et al., 2025). As a legal and design research team based at the University of Edinburgh, our role has been to critically explore the promise and implications of these technologies, with creative professionals and the general public, primarily in a UK context.

[86] https://c2pa.org/
[87] https://decade.ac.uk/

To explore these technologies, we have employed various design research methods (Gaver, 2012; Gaver, Krogh, Boucher, and Chatting, 2022) where we seek to do research through the creative practice of designing and deploying functional, interactive prototypes, drawing inspiration from prior work on 'Technology Probes' (Hutchinson et al., ????) and 'Research Products' (Odom, Wakkary, Lim, Desjardins, Hengeveld, and Banks, 2016), presented to various audiences through as interactive experience and public exhibitions. Design research is a critical approach which enables us to be hands on in exploring how these emerging technical concepts and tools can be applied in practice. By making engaging and tangible artefacts, we are able to engage diverse professionals, communities and publics to explore their understanding and attitudes to these tools and the issues they raise.

Design research is also about looking beyond single solutions or implementation. Instead we are implementing the ORA framework to chart the design space and expose the challenges and requirements of technologies in this context. In this paper we reflect particularly on two iterative research products: the ORAgen Demonstrator and ORAgen Fables, which are each briefly described below [See Figures]. We have demonstrated and deployed both artefacts in various contexts, including a workshop with key UK creative technology and media companies (Liddell et al., 2024b), and other discovery workshops conducted to explore provenance data specifically (Moruzzi et al., 2025), in total we have directly engaged 86 professionals working in the creative industries and over 128 members of the general public, including 33 in-depth interviews.

Prior and concurrent publications describe our research methods and data collection in more detail. In general, our work adopts a qualitative approach, where we use design artefacts as prompts for open-ended interviews and discussion. As a practice-led programme of design research, this work is necessarily emergent (Gaver et al., 2022), where our goals, interest and focus has evolved over time, to thoroughly map and explore different facets of the design space related to ORA and other tokenised licensing approaches.

In this paper, we hence take the opportunity to take stock and reflect on these various engagements, with a view to setting out an interdisciplinary research agenda. Specifically, we ask: how might emerging approaches combining tokenised licensing and content authenticity provide tools that facilitate trust and value exchange between creative rightsholders and AI developers, towards a sustainable ecosystem for digital creative media?

We now briefly introduce our two primary design artefacts based on the ORA Framework before reflecting on the key issues which participants have expressed through their engagements.

3.2 The ORAgen Demonstrator

ORAgen is a web application that demonstrates the ORA framework by enabling users to create, tokenize, license, and remix simple color collages [88]. ORAgen has six predefined demo accounts, each with an associated Ethereum wallet and license contract. After selecting an account, users have the option to either create, reuse, or engage with the tokenised color collages. We designed the ORAgen demonstrator as a technology probe (Hutchinson et al., ????), where ORAgen is used to stimulate discussion with users on the potential opportunities and implications of using a tokenised framework in their own day-to-day activities. We have used the demonstrator as part of an in-depth interview study with 21 different creative professionals including animation, sound, visual arts, digital art, and cultural institutions . Additionally, we have used ORAgen to engage with over 300 creative professionals at industry conferences and events, and 69 creative professionals through two in person workshops and one online workshop (see Liddell et al. (2024a) for further description).

3.3 ORAgen Fables

ORAgen Fables was designed as an iteration of the ORAgen demonstrator for public audiences to experience ORA through a simple walk-up interaction. ORAgen Fables is a collaborative story writing experience based around a central story, image, and up to three characters. Users are invited to add contributions to shape what happens next in the story by adding up to a 240-character contribution described as a 'Fablebit'. At the end of each deployment, these Fablebits were handed over to a 'narrator', a professional storyteller, who created a final version of the story using these crowdsourced contributions. In interacting with ORAgen Fables, each Fablebit is tokenised with users applying customised licenses that provide control over what the narrator can and cannot do with their contribution in the final

[88] Publicly available at the following link: https://oragen.designinformatics.org/

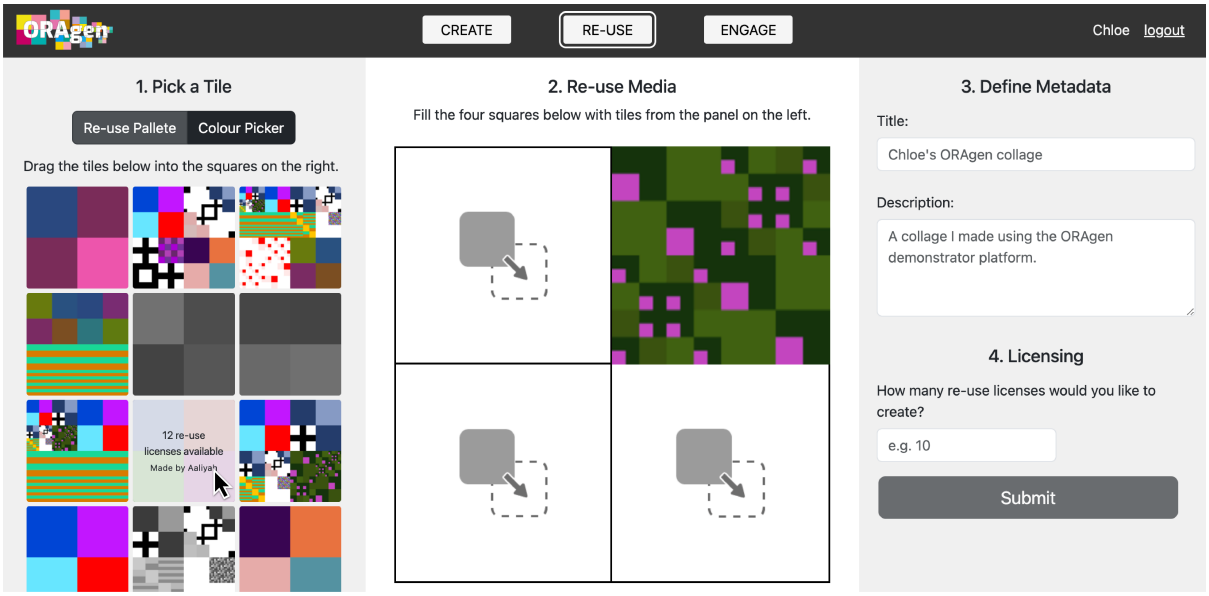


Fig. C1: Screenshot of the ORAgen demonnstrator web app, showing how users can create and reuse simple tiled collages to generate complex patterns, and embed specific metadata and licensing conditions

Fig. C2: Deployments of ORAgen Fables in various public contexts in Edinburgh

story, these included whether their Fablebit could be used to make an AI story, whether they should be attributed, and if it could be adapted. ORAgen Fables was deployed in four locations around Edinburgh, UK, resulting in 128 contributions made and 12 in-depth interviews with users who contributed to ORAgen Fables at one of these deployments.

4 Emergent Issues and Opportunities

4.1 Recognising the Diversity of the Creative Industries and Sector Specific Needs

In 1998, the UK government was one of the first to comprehensively map and define the 'creative industries' as a distinct and significant area of the economy, encompassing such diverse activities as advertising, crafts, design, film-making, publishing, cultural heritage, music, as well as IT, software and computer services (Bakhshi, Freeman, and Higgs, 2013). The precise definitions and mapping of each of these sub-sectors remains an ongoing debate. Each of these sectors fundamentally depend on creative labour and talent, which are employed to produce novel and original things of value to the wider economy and society, and whose outputs are not easily determined in advance (Bakhshi et al., 2013). However, the actual creative practices, norms and traditions, the media and materials involved, the degree of

digital mediation, and the organisation of labour, IP and enterprise in each of these sectors can be vastly different.

In many debates related to copyright and genAI, 'creatives' are often described in a homogeneous way. But while reports from various parts of the sector demonstrate similar sentiments regarding creator rights (Authors Licensing and Collecting Society, 2024; DACS, 2024), this homogeneity in debates risks undermining the impact genAI could have on specific practices, business models, and social norms within the sector. Based on our interviews remixing practices are a case in point, where remixing can play an essential role in creative processes without necessarily the expectation for attribution or compensation, such as on open music platforms like ccMixter [89], or creative movements such as net art (Paul, 2023). While in comedy, the retelling or remixing of another comic's joke is the ultimate faux pas. The conflation of these different sectors could lead to a chilling effect, undermining certain creative practices in a bid to protect others. Similarly, provenance and content authenticity tools are more robust in some media – (such as images and video) – than others (such as text and music), which could lead to assumptions about the level of protection of creative work generally while leaving certain communities vulnerable to their work being scraped and reused without their consent. Protocols such as ORA could enable individual creators and communities to express norms and conditions that reflect the distinctness of their creative practice. However, considering the information ecosystem as a whole, we should be continuously aware of how variably individual creative practices and livelihoods will be affected by GenAI.

4.2 Machine-Readable Granularity to Enable Bespoke and Automated Licensing

Given the scale of GenAI – both in the data that is input, and the media that is output – a key challenge for any solution is to manage a potentially vast number of transactions. Machine readability is hence a key quality. Initial consultations on AI and copyright in the UK have focused heavily on the merits of an opt-out or opt-in system to allow data mining towards AI training. However, rightsholders and others that we have engaged with (DACS, 2024) clearly seek more transparency and granularity in how their work is used by AI systems in order to build trust. Further situated research is required to understand the primary conditions and concerns of rights holders, but these could relate to certain AI models (permitting some, but not others), depend upon the intended output of the model, allow pre-training but not fine-tuning or post-training, or be time-limited. Underpinning these conditions are moral and ethical concerns, financial considerations, as well as practical feasibility of tracking and enacting them.

Machine readability, and granular expression of intent and rights could pave the way for more nuanced and bespoke licensing agreements. Traditional licensing arrangements for digital media tend to be governed by legal contracts and tend to rely upon commercial deals managed by large rightsholders (e.g. publishers, music labels, media organisations etc.), or the terms and conditions imposed by platforms hosting content. In practice, digital rights management (DRM) software is used to enforce and manage content through conditional access to that content. In the context of remix cultures, Creative Commons [90] has become a standard that enables creatives to share their work with flexible permissions on reuse while encouraging creativity (Lessig, 2004). Many individual creatives depend on various intermediaries to manage the complexity of licensing their work – however, there are often then considerable power imbalances and limited trust in how licenses are managed, and limited flexibility available to individual creators or circumstances.

One particular promise of tokenised licensing is that specific rights can be bundled or unbundled into specific tokens. For example, rights to sample parts of the work (but not others), or to train specific AI models could each be represented by different tokens. The potential templates and programmability of these licenses is vast, and could even be highly dynamic or collectively oriented – for example, containing time-based conditions (i.e. limiting when or for how long content can be used), ensuring a split of royalties from derivative works, or issuing only a limited number of licenses. In practice, like Creative Commons, and as has been observed in the development of token standards in the NFT ecosystem, we imagine that various common license templates would be available for creators to choose from when they register and share their work. It is also plausible that smart contracts could be used to configure forms of collective licensing, where rights and royalties are shared directly across communities. For example, our conversations with creative professionals pointed to the potential for smart contracts to support community-based and collaboratively produced work by managing shared control over the remuneration

[89] https://ccmixter.org/
[90] https://creativecommons.org/

through distributing compensation automatically. This has also been demonstrated in exhibitions such as Trickle Down: A New Vertical Sovereignty by Helen Knowles, and ECHOES by EPOCH gallery and Los Angeles County Museum of Art, which are experiments in how smart contracts can be used to support equity-based financial models in collaborations (Liddell, 2024). Being implemented in a decentralised way, there are opportunities for creators to be able to set licenses that are tailored to particular contexts. On paper, especially for individual creators and artists, it could be empowering to have this much individual or community-based control over how their work is licensed.

However, we should remain wary of the core limitations of smart contracts, which are 'book smart' rather than 'street smart' (Levy, 2017), and highly rigid, fixed and immutable forms of licensing. Creators must be highly confident and have a clear understanding of the implications of the license tokens they choose to provide. It may be that tokenised licensing is best suited for highly specific kinds of transactions or permissions (e.g. allowing bots or AI agents to read the data) and not others. Furthermore, as we have seen in NFT marketplaces, despite promises of decentralisation, and by stepping traditional gatekeepers and institutions, in many cases, creators are required to interact with and trust new intermediaries, to support them in engaging technically and socially with these new ecosystems. In practice, any DLT-based systems require considerable literacy and onboarding for new users.

4.3 Tracking Reuse and Attribution

The value of provenance data and tokenised licensing is only realised when working in conjunction with tools that can check and act upon this data. For example, a web browser plug-in could read images on a web page and identify those with embedded provenance data, or else look-up this data where it is absent but can be traced after inspecting the image. Likewise, a publisher or archive could identify the licensing terms attached to an image that is submitted to a platform, and require evidence that the work is original, or that any licensing terms set by the original creator(s) have been met. The success of such systems depends considerably on the interoperability of open standards such as C2PA (Balan et al., 2025; Bennett et al., 2025), security, and incentives for publishers and platforms to seek and require verifiable provenance for media they publish.

Truly interoperable solutions for content provenance and licensing require coordinated standardisation across multiple complementary components – not only for embedding authenticity metadata (e.g. C2PA), but also for content identification (e.g. perceptual descriptors or watermarking), machine-readable rights declarations (e.g. ODRL), and automated licensing and compensation logic. Banks of standardised, vetted licensing contracts could further reduce friction for creative practitioners. Due to the susceptibility of metadata to stripping by content platforms, means to store and look-up (via content identification) missing provenance and licensing information are required for resilience (Collomosse and Parsons, 2024). This implies a governance structure over some trusted registry. One mechanism could be a centralised database run by a trusted authority or vendor. However, standalone or proprietary solutions risk limited interoperability and fragmented adoption (Balan et al., 2025). Decentralised registry infrastructure could support look-up, without recourse to trusted authorities. Yet decentralised registries are challenging to deploy at scale using current DLTs. Solutions are therefore emerging in which multiple vendors operate centralised registries that can federate search via a common interface for decentralised querying. Such a decentralised look-up protocol was standardised within the most recent (version 2.2) release of the C2PA specification, enabling stakeholders to develop tools that can reliably discover and act upon provenance and rights information at scale.

Creative and cultural professionals identified various potential benefits to such systems in our engagements with them. For some, decentralised registries may offer modern means to independently prove and timestamp creative work – much in the same way copies of media contained in a sealed envelope, posted by recorded delivery, could provide a third-party timestamp. In the context of an increasingly synthetic media landscape, having an authentic record of creative work is important. While monetisation clearly matters, an artist's moral rights are also of considerable importance (Kheria, 2019) and underpin trust – especially for those handling cultural collections. These rights include accurate attribution, objecting to false attribution, and 'derogatory treatment' of a work. Tokenised licensing systems such as ORA can clearly be used to preserve attribution, through the registry of a work, and being able to trace the presence or influence of the work in subsequent versions. Bespoke licenses may also serve to indicate beforehand permitted and unpermitted uses or treatment of the work. To avoid the potential for false attribution, a trustworthy underpinning identity system is important (to prevent creators providing false identity or credentials in registering works they did not create or hold copyright for).

Finally, some of those who work extensively within the cultural sector were clear about the extensive labour required to undertake attribution and managing copyright, especially when bringing multiple works together in a collection or exhibition, or in reproducing digital versions of works to be presented in an online collection. In this respect, the prospect of any technology which could automate, streamline, or simplify the labour required would be welcome. However, they also noted that what one can do with collections of licensed works will often depend on the lowest common denominator of the most restrictive license (e.g. a film that shows archival footage with restrictive licensing might not be allowed to be cropped and reused for promotional materials). These license logics form through content supply chains, creating cascaded IP that becomes more difficult to manage as the chain grows (see also Bennett et al. (2025)).

Hence, while bespoke licensing may empower original creators, it raises questions about how subsequent works, inheriting conditions of multiple, bespoke licenses would feasibly work together. For example, if an original creator indicates their work should not be used to train an AI, but a subsequent remix permits this, which takes precedent? As in ORA, each subsequent work produces their own licenses; should these be required to inherit the licenses of prior works, and if so, would licenses simply become increasingly restrictive?

4.4 New Business Models and Sustainable Content Supply-Chains for AI

In a survey of UK creatives, non-profit and collective management agency DACS found that 84% of creators would opt-in to license their works for AI usage if they were fairly compensated (DACS, 2024). While there are clearly many significant concerns among artists and creators regarding current commercial AI systems, this survey suggests that fair consent, credit, and compensation could develop better relations and trust between creatives and these technologies. Indeed, it is important to recognise that many artists themselves are already working critically to train and explore their own AI models (Bennett et al., 2025), highlighting a willingness to work with these technologies.

From the perspective of anyone developing or employing AI models, their value and sustainability depends on a trustworthy information ecosystem to supply the data, content, and IP that is required. As noted in the introduction, we should ask ourselves what conditions, systems, and regulations would best serve and incentivise the maintenance of a healthy information ecosystem and hence, a sustainable data supply chain for AI? For artists, beyond the 3 C's (consent, credit and compensation), we know that more transparency about how their work is used is important, as is supporting creative communities themselves in understanding and critically engaging with AI tools (Authors Licensing and Collecting Society, 2024; Moruzzi et al., 2025; Sichani et al., 2025). For someone developing, fine-tuning or prompting an AI model, access to a diversity of well-labelled, secure, and trusted content is important. This is likely to become a growing challenge, as access to quality and easily accessible public datasets decline (and is evidenced in specific deals that AI companies have made with publishers and rightsholders to access paywalled content).

Hence, in terms of bespoke licensing enabled through ORA, we suggest the potential for creators to more intentionally create work that contributes to an AI workflow. Such work could be clearly labelled and licensed, giving developers and users confidence to trust the authenticity of work via its provenance. Rather than supporting general training of foundation models, this approach may be more appropriate where specific datasets or content are used in fine-tuning or inference to aid the personalisation of model training for creatives using small data sets in their practice, or trace the history of iterations an a creative's vision as they co-create using an AI model (Huang et al., 2025).

The specific commercial deals that AI companies have made with large rightsholders partly mirror how music streaming companies dealt with major record labels following the emergence of illegal torrent / streaming services such as Napster (David, 2016). Though there is now unparalleled access to music, many have noted how poorly this business model compensates individual artists. We should therefore carefully consider the potential marketplaces that would serve independent creators and existing creative and cultural institutions with potentially rich datasets which could derived from their work and collections (Terras, Coleman, Drost, Elsden, Helgason, Lechelt, Osborne, Panneels, Pegado, Schafer, Smyth, Thornton, and Speed, 2021), and where there are opportunities for genuine value co-creation in the creative sector. For example, marketplaces might support independent creators to access, remix, or license content for derivative works, while institutions benefit from licensing revenue. Platforms like Hugging Face may support markets for creators to offer fine-tuned models trained on their own unique datasets, or access datasets with specific licensing conditions. Experimental Creative Commons-aligned

infrastructures for co-authorship and collaborative creative production could enable using datasets as shared, public resources.

4.5 Enabling Collective Governance and Representation

A key issue in the debates between creative rightsholders and AI companies is the diversity of actors and distribution of power. AI developers themselves may be artists; creative rightsholders vary from amateurs and poorly represented independent freelancers, to major rightsholders and publishers – all of whom may have vastly different perspectives on fairness, trust, and value in a creative ecosystem. In addition, in the context of AI training taking place on many millions of data points, any efforts to attribute, credit, and compensate individual artists or works can rapidly become implausible (Bennett et al., 2025). Traditionally in the creative industries, forms of collective representation and governance have been necessary to advocate or work on behalf of individuals (e.g. collecting societies, trade unions) with varying degrees of success and fairness (Band and Butler, 2013). Hence, what forms of collaboration and collective governance and representation may be valuable for rightsholders in regard to GenAI?

DLTs often imply individualised, libertarian marketplaces, however by providing infrastructure for peer-to-peer transactions, DLTs can also support new forms of collaboration and distributed governance, for example through DAOs (Decentralised Autonomous Organisations) (Hassan and De Filippi, 2021). As highlighted in Section 4.2, our conversations with creative professionals have hinted at the potential of using tokenised licensing for community-oriented compensation. However, the additional governance layer implemented using a DAO could create a type of data trust where individuals can state their intentions for data use and the DAO acts as the trustee with responsibility for stewarding data on behalf of these individuals (Nabben, 2022). DAOs are built on smart contracts and tokens that enable individuals to vote and automate decision-making. How automated this process is can vary, where on the one hand DAOs may be fully governed through automation, facilitated by humans, or on the other hand be governed by humans facilitated by smart contracts (De Filippi and Wright, 2018). Whichever approach used, DAOs support a shared governance model for data that can reflect the wants and needs of a community of rights holders. In this situation, the smart contracts associated with employing ORA could be managed by the DAO, enabling that community to decide on what specific licensing to apply to data sets with a royalty system that links back to the DAO's smart contracts.

4.6 Limitations

Our research fundamentally has been to explore the extent to which the combination of tokenised licensing and content authenticity tools could be implemented to increase trust and value between creative rightsholders and AI and the wider creative ecosystem in an age of GenAI. There are some clear limitations and challenges to acknowledge with respect to these core technologies and their application in this context.

Employing a form of DLT-based system to assert, register, and record the ownership and rights associated with media assets is attractive to enable interoperability across borders and trustless actors. Just as standard token protocols enabled diverse innovation and experimentation with NFTs, we see ORA, and other tokenised licensing frameworks not as a single solution, but infrastructural layer which can support particular forms of data and value exchange. However, inherently DLTs also produce new governance and usability challenges, and hence may depend upon new intermediaries, for better and worse. Though it has not been a detailed feature of our various deployments and studies, a trustworthy and privacy-preserving identity system is also required. In addition, further research and care should be taken to consider exactly what provenance data and representations of value can and should be reified through tokenisation. Just as we should be wary of a 'copyright trap' (Craig, 2024), we should be cautious to view tokenising and licensing all expressions of creativity as a desirable end-point.

As described previously, C2PA and ORA are not a form of DRM system, in that they do not seek to control or condition access to content. Instead they independently offer standard and persistent means to express ownership and rights related to individual media, which other systems can access and act upon. The ACCCT report (Bennett et al., 2025) offers several examples of different possible marketplaces and applications which might then be enacted upon this registry. Yet, in mapping this space, Balan and Collomosse note that currently no single end-to-end integration of all necessary components (a full 'Content ARC' (Balan et al., 2025)) has been demonstrated and continued open standards development is essential to ensure successful integration of all necessary technical components. Our research has also

not so far considered how tokenised licensing could be applied to historical work or alongside existing media archives.

5 Future Research Questions for Tokenised Licensing and GenAI

Technology and legislation is clearly rapidly advancing in this context. In the meantime, numerous reports, frameworks and proposals such as ACCCT (Bennett et al., 2025), CC Signals(Hardinges, Pearson, and Ross, 2025) and others (e.g. Kretschmer, Meletti, Bently, Cifrodelli, Eben, Erickson, Iramina, Li, McDonagh, Perot, Porangaba, and Thomas (????); Sichani et al. (2025)) are considering possible technological and regulatory solutions to address the issues arising from a lack of transparency, trust, and fair value exchange between creative rightsholders and AI companies. In our research we have sought to surface and understand end-user perspectives on tokenised licensing and content authenticity as one potential remedy. As an interdisciplinary research team, fundamentally grounded in a socio-technical view of innovation, considering how these technologies can support the conditions likely to produce a vibrant and sustainable creative media infrastructure and equitable digital economy has been foundational to our approach. Considering these aims, we conclude with a series of provocations and research questions for further critical research, regarding specific technologies, the regulation they support or require, and their resultant use.

Accounting for Diverse Creative Industries

- How do the needs and effectiveness of content authenticity and tokenised licensing work across different media and industries, especially for non-visual media?
- What specific kinds of license templates or rights expressions would creators in different industries wish to set in relation to GenAI?
- How could existing collecting societies and rights management systems in different fields employ or benefit from tokenised licensing systems?
- How can distinct community norms and creative traditions related to remix, collaboration and copyright be respected through these technologies?

Provenance Data Interactions

- What forms of provenance data are easiest to record, and most valuable to share throughout an AI content supply chain? What kinds of provenance data are difficult to record or likely to be overlooked?
- How might we surface and share relevant provenance data with end-users in meaningful ways?
- What do AI developers most value or require in terms of provenance data and licensing to produce responsible and trustworthy AI systems?
- Are there forms of provenance data, or aspects of the creative process which should not be surveilled and recorded?
- How can and should attribution be recorded in relation to training data, where many millions of data points inform a single output?

Governance, Regulation and Markets

- To what extent should tokenised licensing be used as a tool to enforce copyright law versus a tool to empower creatives beyond what copyright can or should protect?
- How well do tokenised licensing systems function over time? Can they be made more adaptive to changes in the real world?
- To what extent do decentralised approaches atomise and burden individuals more? Or can they support trusted collective and community approaches for wider public benefit? Are these approaches best for individual or collective licensing?
- What further practical support, skills, tools and intermediaries would artists and creators require to engage with these technologies to protect their work and express rights and intentions?
- What new and equitable business models and marketplaces could be built upon these frameworks and tools if they enabled trust between new actors?

These questions can only be addressed through interdisciplinary perspectives and methods, including core research with end-users, further technical development, deployments of prototypes in real-world contexts, and studies of the regulatory and commercial landscape, in various jurisdictions. Through our design-led approach, we hope to have demonstrated the value in being able to materialise and critically explore what provenance backed tools such as ORA and C2PA may be employed to achieve.

Acknowledgements. We thank all of our participants, both professionals and members of the general public, for their generous engagement and participation in our research. This research was supported by the UKRI/EPSRC Next Stage Digital Economy Centre in the Decentralised Digital Economy (DECaDE) (EP/T022485/1), and we are grateful for the support of the wider DECaDE team.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Authors Licensing and Collecting Society (2024) A brave new world? a survey of writers on ai, remuneration, transparency and choice. Tech. rep., ALCS, URL https://d16dqzv7ay57st.cloudfront.net/uploads/2024/12/A-Brave-New-World-_ALCS_AI_Report.pdf

Bakhshi H, Freeman A, Higgs P (2013) A dynamic mapping of the uk's creative industries. Tech. rep., Nesta, URL https://www.nesta.org.uk/report/a-dynamic-mapping-of-the-uks-creative-industries/

Balan K, Agarwal S, Jenni S, et al (2023a) Ekila: Synthetic media provenance and attribution for generative art. https://doi.org/10.48550/arXiv.2304.04639, proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 913-922

Balan K, Gilbert A, Black A, et al (2023b) DECORAIT: Decentralized opt-in/out registry for ai training. In: Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production, pp 1–10, https://doi.org/10.1145/3626495.3626506

Balan K, Gilbert A, Collomosse J (2025) Content ARCs: Decentralized content rights in the age of generative ai. https://doi.org/10.48550/arXiv.2503.14519

Band J, Butler B (2013) Some cautionary tales about collective licensing. Michigan State International Law Review 687

Bennett J, Collomosse J, Gregory-Clarke R, et al (2025) Time to accct: Providing creative industries and ai developers with a copyright framework of access, control, consent, compensation and transparency. URL https://www.costarnetwork.co.uk/resources/time-to-accct

Brook O, O'Brien D, Taylor M (2020) Culture is bad for you:. Manchester University Press, Manchester, England, 10.7765/9781526152152, URL https://www.manchesterhive.com/view/9781526152152/9781526152152.xml

Brown I (2023) Allocating accountability in ai supply chains. Tech. rep., Ada Lovelace Institute, URL https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

Collomosse J, Parsons A (2024) To authenticity, and beyond! building safe and fair generative ai upon the three pillars of provenance. IEEE Computer Graphics and Applications 44(3):82–90. https://doi.org/10.1109/MCG.2024.3380168

Craig C (2024) The copyright trap. Osgoode Legal Studies Research Paper No 4905118 URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4905118

DACS (2024) Artifical intelligence and artists' work. URL https://cdn.dacs.org.uk/uploads/documents/News/Artificial-Intelligence-and-Artists-Work-DACS.pdf

Daly S, Hearn G, Papageorgiou K (2025) Sensemaking with ai: How trust influences human-ai collaboration in health and creative industries. Social Sciences & Humanities Open 11:2590–2911. https://doi.org/10.1016/j.ssaho.2025.101346

David M (2016) The legacy of napster. In: Networked Music Cultures: Contemporary Approaches, Emerging Issues. Springer, p 49–65

De Filippi P, Wright A (2018) Blockchain and the Law. Harvard University Press

Duffy BE, Poell T, Nieborg DB (2019') Platform practices in the cultural industries: Creativity, labor, and citizenship. Social Media + Society 5(4). https://doi.org/10.1177/2056305119879672

Gaver W (2012) What should we expect from research through design? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '12, p 937–946, URL https://doi.org/10.1145/2207676.2208538

Gaver W, Krogh PG, Boucher A, et al (2022) Emergence as a feature of practice-based design research. In: Proceedings of the 2022 ACM Designing Interactive Systems Conference. Association for Computing Machinery, New York, NY, USA, DIS '22, p 517–526, URL https://doi.org/10.1145/3532106.3533524

Government U (2025) Ai opportunities action plan. Tech. rep., Department for Science, Innovation & Technology, URL https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

Hardinges J, Pearson S, Ross R (2025) From human content to machine data; introducing cc signals. Tech. rep., Creative Commons, URL https://creativecommons.org/2025/06/25/introducing-cc-signals-a-new-social-contract-for-the-age-of-ai/

Hassan S, De Filippi P (2021) Decentralized autonomous organization. Internet Policy Review https://policyreview.info/glossary/DAO

Huang J, Yew RJ, Venkatasubramanian S (2025) Copyrighting generative ai co-creations. In: Proceedings of the 2025 ACM Designing Interactive Systems Conference. Association for Computing Machinery, New York, NY, USA, DIS '25, p 1156–1164, 10.1145/3715336.3735683

Huang S, Siddarth D (2023) Generative ai and digital commons. arxiv https://doi.org/10.48550/arXiv.2303.11074

Hutchinson H, Mackay W, Westerlund Bea (????) Technology probes: Inspiring design for and with families. https://doi.org/10.1145/642611.642616

Keller P (2025) Beyond ai and copyright. Tech. rep., Open Future, URL https://openfuture.eu/publication/beyond-ai-and-copyright/

Kheria S (2019) Copyright in the everyday practice of writers. In: Jefferies J, Kember S (eds) Whose Book is it Anyway? Open Book Publishers, p 141–180, https://doi.org/10.11647/OBP.0159.07

Kretschmer M, Meletti B, Bently L, et al (????) Copyright and AI: Response by the CREATe centre to the UK government's consultation. 10.5281/zenodo.14931964, URL https://doi.org/10.5281/zenodo.14931964

Lessig L (2004) Free Culture. The Penguin Press

Levy KE (2017) Book-smart, not street-smart: blockchain-based smart contracts and the social workings of law. Engaging Science, Technology, and Society 3:1–15. https://doi.org/10.17351/ests2017.107

Liddell F (2024) The distributed museum: Using smart contracts and NFTs to find financial equity in co-produced exhibitions. In: Financial Management in Museums. Routledge, p 16, https://doi-org.eux.idm.oclc.org/10.4324/9781003308003

Liddell F, Tallyn E, Morgan Eea (2024a) Oragen: Exploring the design of attribution through media tokenisation. https://doi.org/10.1145/3656156.3663693

Liddell F, Velasco N, Schreeve C, et al (2024b) Oragen: Emerging futures for tokenisation and digital media rights. URL https://www.digicatapult.org.uk/publications/post/oragen-emerging-futures-for-tokenisation-and-digital-media-rights/

Marshall L (2015) Let's keep music special. f—spotify: on-demand streaming and the controversy over artist royalties. Creative Industries Journal 8(2):177–189. https://doi.org/10.1080/17510694.2015.1096618

Moruzzi C, Tallyn E, Liddell F, et al (2025) Content authenticities: A discussion on the values of provenance data for creatives and their audiences. In: Proceedings of the 2025 Conference on Creativity and Cognition. Association for Computing Machinery, New York, NY, USA, C&C '25, p 128–140, URL https://doi.org/10.1145/3698061.3726918

Nabben K (2022) Decentralised autonomous organisations (daos) as data trusts: A general-purpose data governance framework for decentralised data ownership, storage, and utilisation. SSRN http://dx.doi.org/10.2139/ssrn.4009205

Nieborg DB, Poell T (2018) The platformization of cultural production: Theorizing the contingent cultural commodity. New Media & Society 20(11). https://doi.org/10.1177/1461444818769694

Noh M, Odenkirk S, Shionoiri Y (2023) GM! time to wake up and address copyright and other legal issues impacting visual art nfts. Columbia Journal of Law & the Arts 45(3)

Odom W, Wakkary R, Lim Yk, et al (2016) From research prototype to research product. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '16, p 2549–2561, URL https://doi.org/10.1145/2858036.2858447

Paul C (2023) Digital art. Thames & Hudson

Shumailov I, Shumaylov Z, Zhao Y, et al (2024) Ai models collapse when trained on recursively generated data. Nature 631(8022):755–759. URL https://www.nature.com/articles/s41586-024-07566-y

Sichani A, Westenberger P, Bryan-Kinns N, et al (2025) Braid researchers' response to uk government copyright and ai consultation. https://zenodo.org/records/14945987

Terras M, Coleman S, Drost S, et al (2021) The value of mass-digitised cultural heritage content in creative contexts. Big Data & Society 8(1):20539517211006165. 10.1177/20539517211006165, URL https://doi.org/10.1177/20539517211006165

From Black Box to Glass Box: Leveraging Blockchain to Audit AI Systems Through Multistakeholder Participation

Roberto Caparroz de Almeida[1], Lucas Maldonado Latini[2]

[1] Professor, FGV Direito SP, São Paulo, Brazil. Visiting Professor at MIT (2024).
[2] Master in Law and Technology, FGV Direito SP, São Paulo, Brazil. Visiting Researcher, DCU Law & Tech Research Cluster.

Corresponding author. E-mail: roberto.almeida@fgv.br

Abstract

A pervasive trust deficit plagues AI's widespread adoption due to its "black box" nature, an opacity that makes it challenging to understand how certain outputs or decisions are reached. Previous studies have looked at the use of blockchain technology for AI audits. However, they lack the knowledge to make a multistakeholder approach feasible. To address these shortcomings, our research proposes a framework that harnesses the power of blockchain technology for AI auditing by including several stakeholders. Blockchain's core attributes of decentralisation, immutability, and cryptographic security offer the opportunity to record and verify AI processes transparently. Specifically, we suggest the implementation of a blockchain-based audit trail which stores cryptographic proofs detailing model parameters, data sources, and algorithmic updates at various checkpoints that different actors could verify. This shared verification process democratises oversight, mitigating the concentration of control in the hands of a single authority or company, or even a few local authorities or companies. Taken together, our analysis argues that blockchain auditing is a crucial strategy for transforming AI from an opaque, centralised "black box" into a collaborative, verifiable "glass box". A central advantage of our proposal is its capacity to foster multistakeholder governance. In a decentralised ledger environment, diverse participants, including regulators, academic institutions, professional bodies, and civil society representatives, can collectively validate AI systems' integrity.

Keywords

Artificial Intelligence, Blockchain, Framework, Auditing.

1 Introduction & Problem Statement

The widespread use of AI in sensitive sectors such as healthcare and public services has created an unprecedented crisis of verifiability. This is because the way AI models process data and respond to user interactions lacks transparency, leading to the well-known term "black box" (Pasquale, 2016).

This opacity is particularly serious in multilingual and regional contexts. The main generative AI models are primarily trained in English, and the data used, derived from automatic internet scans, does not necessarily reflect the cultural or social nuances of non-English speakers (Caparroz, 2024). This linguistic and cultural disparity is more than just a technical problem or an issue with data origin; it reveals a profound ethical and social challenge.

The predominance of English in AI model training creates a significant asymmetry. This is because the most relevant content in other languages, such as books and scientific articles in Portuguese, for example, is often only available behind paywalls. This scenario raises economic and legal issues, including copyright concerns.

Furthermore, given that most available benchmarks are exclusively in English, it is difficult to accurately assess the quality of responses and interactions in other languages.

In the interaction between users and generative AI models, it is common to observe what we term the "logic of wonder" (Caparroz, 2024). This refers to a purely subjective perception that the responses presented are excellent and perfectly aligned with individual interests. However, this is not always the case, as the quality of interactions in Portuguese and other less widely spoken languages tends to be lower than those produced in English, even when both refer to the same subject.

The methods currently available for evaluating AI models, such as the benchmarks, model cards, algorithmic impact assessments, and third-party audits, are insufficient and often disregard regional context. These approaches prevent any form of participatory collaboration, thereby excluding researchers, academics, and civil society organisations from the process of assessing and validating the impact of AI on people's daily lives.

This article proposes a new audit framework, leveraging blockchain technology, to transform the current centralised paradigm into a distributed, transparent, and verifiable analysis model. The intrinsic attributes of blockchain - such as immutability, decentralisation, and cryptographic security - address the limitations of traditional audits, enabling a more democratic governance model for applications involving AI.

The availability of an international, open-source platform for evaluating AI models allows public and private entities from different countries to create their own contextual evaluation parameters. Records stored on the blockchain make it possible to identify those responsible for the evaluations and build a repository of audited data for traceability, provenance, and regionalised analysis. Based on these auditable logs, regulators, the academic community, and other stakeholders will be able to discuss and, in the future, implement standards aimed at addressing local and specific problems, based on objective and independently verifiable criteria.

We propose the creation of a proof of concept based on Hyperledger Fabric, an open-source blockchain network widely used by the international community. Our evaluation model allows for the development of a collaborative environment where records are fully identifiable and traceable, with low latency and cryptographic security, and capable of operating on a large scale.

The remainder of this article is structured as follows: Section 2 provides an overview of the concepts of auditing in AI models, their problems, and limitations. Section 3 presents a hypothetical framework for auditing AI models using blockchain technology. Section 4 presents an evaluation of the framework from a practical point of view. Finally, section 5 presents the conclusion and opens space for future work.

2 Background & Related Work

2.1 Core audit concepts and their application to the AI context

Auditing, whether in financial, institutional, or computational contexts, relies on foundational principles that guarantee reliability, traceability, and trust (Habibikilak, 2024). These include transparency, accountability, and provenance, which together form the backbone of any rigorous governance framework, including in the AI field (Targino, Souza, and Salgado, 2024).

Transparency is regarded as a prerequisite for legitimacy and public trust. It is closely related to the degree to which an observer can access and comprehend information about a system or decision-making process. It is not only about data disclosure or mere access to information but also about the interpretability, context, and meaningfulness of that information to different stakeholders (Kiseleva, Kotzinos, and De Hert, 2022).

In the AI context, transparency relates to understanding the internal workings of models, system design decisions, and how data is used during training and inference (Liao, Subramonyam, Wang, and Wortman Vaughan, 2023). However, AI systems are frequently described as "black boxes" due to the opacity of complex models, such as deep neural networks. This lack of transparency hinders stakeholders, including users, regulators, and affected communities, from understanding or challenging decisions. Tools such as explainable AI and model documentation aim to enhance transparency (Casper, Ezell, Siegmann, Kolt, Curtis, Bucknall et al., 2024); however, they remain insufficient for lay understanding, democratic scrutiny, and accountability purposes (Busuioc, Curtin, and Almada, 2023).

Accountability refers to the obligation of individuals or institutions to explain, justify, and take responsibility for their actions (Williams, Cloete, Cobbe, Cottrill, Edwards, Markovic, Naja, Ryan, Singh, and Pang, 2022). It involves clearly defined roles, mechanisms for enforcement (Williams et al., 2022), and channels for redress. In the context of personal data, accountability is a core principle introduced in 2016 by the General Data Protection Regulation (GDPR), where data controllers are responsible for, and must be able to demonstrate compliance with, several other principles, including data minimisation, accuracy, and transparency.

In the realm of AI, accountability is often weakened by the distributed and opaque nature of machine learning pipelines. Although recent EU legislation attempts to address this issue, scholars argue that determining who is responsible for AI errors (e.g., data engineers, developers, deployers, or users) remains unclear (Kempton, Parmiggiani, and Vassilakopoulou, 2023). Therefore, formal and robust audit mechanisms are increasingly seen as necessary to identify responsible actors, track oversight gaps, and prevent systemic evasion of liability.

Lastly, provenance refers to the history, origin, and flow of information or objects. In digital systems, provenance is the origin of data and the sequence of computational steps and transformations it undergoes (Pérez, Rubio, and Sáenz-Adán, 2018). It captures how data is derived, transformed, and reproduced, forming the basis for data reliability, reproducibility, and auditability (Aldeco-Pérez and Moreau, 2010).

Provenance is vital to the auditability of AI systems, ensuring traceability of data and model changes. This is particularly relevant for legal compliance, scientific reproducibility, and trust-building, as it allows auditors to identify when and how models were altered or retrained (Targino et al., 2024).

Together, these three principles create a foundation upon which reliable AI auditing and governance can be constructed, ensuring functional correctness, democratic legitimacy, and societal trust.

2.2 Limitations of current AI-audit toolchains

AI audit tools, such as model cards, black-box audits, and third-party audits, represent early yet incomplete responses to the growing demand for transparency, accountability, and provenance in algorithmic systems. Despite their great intentions, these tools frequently fail to provide rigorous, enforceable, and stakeholder-inclusive accountability.

Model cards aim to document an AI model's intended use, performance metrics, potential ethical considerations, and other features to inform stakeholders about the system's characteristics (Mitchell, Wu, Zaldivar, Barnes, Vasserman, Hutchinson, Spitzer, Raji, and Gebru, 2019). However, these artefacts often remain too technical for policymakers or impacted users (Bracamonte, Pape, Löbner, and Tronnier, 2023) and lack standardisation across different organisations. While model cards represent a step towards transparency (Shen, Wang, Deng, Brusse, Velgersdijk, and Zhu, 2022), their practical utility is hampered by poor accessibility, inconsistent implementation, and the absence of verification protocols.

Furthermore, research reveals a gap between the proposal and practice regarding documentation practices, such as model cards (Bhat, Coursey, Hu, Li, Nahar, Zhou, Kästner, and Guo, 2023), as they do not inherently include mechanisms for accountability and do not ensure the traceability of decisions over time. Without verifiable provenance or immutable logs, these tools provide only a snapshot of the model's design based on the information given by AI providers or deployers themselves, thus remaining insufficient for retrospective audits or stakeholder challenges.

When it comes to black-box audits, such as algorithmic impact assessments (AIA), several initiatives have already been proposed by different stakeholders (Reisman, Schultz, Crawford, and Whittaker, 2018) (Ada Lovelace Institute, 2022) (Treasury Board of Canada Secretariat, 2024). These tools usually suggest methodologies for organisations to conduct and document their self-assessments regarding the development or deployment of AI systems in many areas.

However, besides having to once again rely on the information provided by organisations, just like the model cards technique, when reviewing the AIA's documentation, auditors and other stakeholders can only test inputs and observe outputs without access to the underlying model, training data, or deployment processes. This extends beyond mere technical limitations, thereby reflecting an information asymmetry that privileges model providers over affected communities and external reviewers, thus undermining proper accountability.

Lastly, there are third-party audits, where organisations hire outsourced services from so-called independent parties to audit their AI systems, models, or processes (Lam, Lange, Blili-Hamelin, Davidovic, Brown, and Hasan, 2024). More recently, due to the rapid advancement of AI, important organisations have issued specialised certifications for AI audit (ISACA, 2025) and governance (International Association of Privacy Professionals (IAPP), 2025) professionals, as well as non-binding standards for AI management systems (ISO, 2023), indicating a growing demand for this kind of service.

Despite independence and transparency being core characteristics of third-party audits, these elements are often undermined in such auditing processes (Raji, Xu, Honigsberg, and Ho, 2022). Many audit reports are not publicly released, and audit methodologies often lack standardisation or regulatory oversight (Lam et al., 2024). These conditions erode public trust and enable AI providers or deployers to influence the narrative surrounding audit outcomes.

Third-party AI audits also have issues with the technical tools used in the auditing process. A landscape analysis of over 400 AI auditing tools revealed that most focus on elements such as model evaluation or bias detection, neglecting the broader needs of audit practitioners, including harm discovery (Ojewale, Steed, Vecchione, Birhane, and Raji, 2025). Interviewed practitioners expressed that many existing tools are disconnected from practical audit workflows, lack integration with legal reporting, and fail to support advocacy or policy use (Ojewale et al., 2025).

Even when audits are well-executed with adequate tools, their impact may be limited by institutional and regulatory constraints. Without a supportive governance ecosystem, including public transparency mandates, mechanisms of redress, and stakeholder participation, audits risk becoming merely performative. As (Birhane, Steed, Ojewale, Vecchione, and Raji, 2024) argue, the current AI audit landscape represents a "broken bus" on the road due to its inability to deliver real accountability without systemic reforms (Birhane et al., 2024).

In conclusion, while current AI audit techniques have introduced foundational practices for transparency and risk mitigation, they fall short in several key dimensions, including the depth of access, independence, stakeholder involvement, and enforceability. Addressing these limitations will require moving beyond fragmented, black-box assessments toward comprehensive, immutable, decentralised, and verifiable audit infrastructures that include all affected parties and ensure verifiability.

2.3 Blockchain fundamentals and audit relevance

As the limitations demonstrate, ensuring verifiable transparency and accountability has become a significant challenge in auditing AI systems. Blockchain, formally introduced and popularised as the infrastructure for Bitcoin (Nakamoto, 2008), offers a promising suite of properties that directly address core auditability needs in AI systems. These properties include immutability, decentralisation, and verifiability, all of which help resolve key weaknesses in current AI audit toolchains.

Immutability refers to the inability to alter stored records, a property achieved through cryptographic hashing and consensus protocols in blockchain systems (Yadav, Singh, and Kushwaha, 2023). This characteristic is vital for creating tamper-proof audit logs, enabling independent verification of the decisions made, when they were made, and by whom (Ahmed, 2025). Once stored on a blockchain, audit records such as model training history, data lineage, or decision explanations become permanent, timestamped, and verifiable (Yadav et al., 2023), offering stronger guarantees than mutable documentation tools like model cards.

In the context of AI auditing, immutability refers to creating audit records that, once written, cannot be altered or deleted by system operators, developers, or external parties (Sutton and Samavi, 2018). A wide range of relevant events in AI development and deployment can be captured by the immutability

property of blockchain technology (Butt, Junejo, Ghulamani, Mahdi, Shah, and Khan, 2023), including model training histories (e.g., hyperparameters, data versions), model updates and redeployments, data provenance records, decisions and outputs for high-stakes use cases, and relevant user interactions. This is critical for ensuring that AI system operations can be independently verified and that records cannot be retroactively manipulated to obscure unethical or illegal behaviour.

Decentralisation is another core blockchain property with direct relevance to AI auditing. In traditional auditing infrastructures, critical audit records, such as model documentation, risk assessments, and performance logs, are typically stored in centralised repositories controlled by a single actor (Ojewale et al., 2025), usually the AI provider or a designated auditor. This centralisation creates major risks such as single points of failure, potential for internal tampering, and power asymmetries between AI developers and external stakeholders (Urman, Smirnov, and Lasser, 2024).

Blockchain systems eliminate these risks by replacing single points of trust with a distributed network of validating nodes. Each node holds a copy of the ledger, and new records are accepted only through consensus protocols. As a result, no single actor can unilaterally alter or erase audit records (Zheng, Xie, Dai, Chen, and Wang, 2017).

In an AI auditing context, decentralisation can support, for example, (i) cross-institutional audits where regulators, civil society groups, and industry bodies participate equally in maintaining the audit ledger, and (ii) immutable multi-party provenance tracking, where model updates, training data versions, and decision outcomes are logged in a way that no single stakeholder can alter post hoc. Decentralisation further supports security and governance by preventing AI providers from monopolising the audit process, enhancing transparency to allow each participant to verify blockchain-stored audit events independently, and improving resilience through majority-verification mechanisms that preserve audit integrity.

The last relevant blockchain property in the auditing context is verifiability, understood as blockchain's support for cryptographic proofs of record authenticity. Every entry in the chain is mathematically verifiable through hash functions and digital signatures (Zheng et al., 2017), thus allowing all participants to independently confirm the information without needing to trust the entity maintaining the register.

In conventional AI governance, even if transparency is nominally provided (e.g., through model cards or audit reports), external parties have limited ability to independently verify whether claims are accurate or whether an audit trail is complete and unaltered (Casper et al., 2024). Blockchain changes this dynamic precisely by offering cryptographically verifiable records that can be mathematically validated. Moreover, digital signatures attached to blockchain transactions provide proof of authorship and integrity (Zheng et al., 2017), ensuring an adequate level of certainty about stakeholder participation.

In the field of AI audit, verifiability combines with immutability and decentralisation, enabling proof of several key steps in the development or deployment phases, such as model lineage, where auditors can verify that the deployed model matches its certified version (Severino, Canciani, Felicioli, Gervasi, Pelosi, Natali, Salvatore, and Russo, 2024). It can also support proof of decision-making in real-time systems, which is especially relevant for high-risk AI systems under the EU AI Act (e.g., healthcare and credit score), since verifiable logs of inferences can provide reliable evidence of how a specific outcome was reached.

In sum, these three properties of blockchain (immutability, decentralisation, and verifiability) offer a robust foundation for addressing the limitations of existing AI audit mechanisms. By enabling tamper-proof, distributed, and independently verifiable audit records, blockchain can significantly enhance the transparency, accountability, and resilience of AI governance frameworks. However, while the theoretical potential is clear, practical implementations of AI and blockchain auditing remain at an early stage. In the next section, we critically examine existing proposals that seek to operationalise these concepts, highlighting both their contributions and the persisting gaps that motivate the development of the more robust, multistakeholder audit framework proposed in this article.

2.4 Review of relevant proposals

Over the past years, scholars have explored the integration of blockchain technologies into auditing, including in AI governance, particularly to improve transparency, accountability, and auditability. These proposals typically leverage blockchain's core properties to create audit trails that are resistant to tampering and accessible to multiple parties. While these efforts represent important steps, many remain conceptual or limited in scope, and substantial challenges persist.

Recently, Akther, Arobee, Adnan, Auyon, Islam, and Akter (2025) proposed a blockchain-based framework to enhance AI transparency by recording the behaviours of AI systems, including decision inputs, model parameters, and performance metrics such as data descriptors. The framework leverages blockchain's immutability and transparency to create an auditable trail of AI decisions, thereby strengthening accountability and fostering public trust. However, the proposal lacks concrete mechanisms for verifying the authenticity of recorded logs and does not support real-time access or inspection by external auditors.

Balan, Learney, and Wood (2025) propose a cryptographic framework for achieving end-to-end verifiability in AI pipelines, aiming to ensure the integrity and traceability of datasets, model artefacts, and outputs throughout the development lifecycle. Their approach involves linking each component through cryptographic hashes and metadata records, enabling tamper-evident provenance chains. While the framework could be supported by blockchain infrastructure, the paper does not implement or formally integrate blockchain into its design; instead, it mentions other initiatives that utilise blockchain. Moreover, it does not address how these verifiability mechanisms could be operationalised in real-world systems or made accessible to non-technical stakeholders.

From a regulatory and legal standpoint, Ramos and Ellul (2024) examine how blockchain can facilitate the EU AI Act's requirements for transparency and accountability. Their proposal focuses on using blockchain to maintain immutable logs of AI operations and to manage data access through smart contracts. This approach strengthens audit integrity and facilitates automated compliance verification. However, the proposal is limited in scope to internal compliance and assumes a closed, enterprise-centric model of auditability. It does not address how independent external actors (e.g., civil society or academic auditors) would access these records or challenge misleading claims.

Other authors propose a more technical blockchain-enabled framework for auditing federated learning processes, emphasising verifiable and decentralised oversight of the training phase (Kalapaaking, Khalil, Yi, Lam, Huang, and Wang, 2025). Their architecture uses smart contracts to record model update submissions, verify conformance to learning objectives, and coordinate aggregation through multisignature schemes and consensus mechanisms. This approach enhances the integrity and traceability of distributed training without requiring centralised control. However, the framework is primarily designed for technical stakeholders within a closed federated learning environment. It does not include mechanisms for post-deployment monitoring, interpretability, or compliance with legal audit requirements. It also does not address how external auditors or regulatory bodies could access or challenge system behaviour.

Similarly, (Potdukhe, Gourshettiwar, Zade, and Waghale, 2025) present a comprehensive review of the integration of blockchain, federated learning, and explainable AI (XAI), proposing that blockchain enhances federated learning systems by ensuring tamper-resistant audit trails and improving trust through transparent logging of model updates and participant contributions (Potdukhe et al., 2025). The authors discuss how smart contracts, consensus mechanisms, and immutable ledgers can be leveraged to improve traceability and accountability within decentralised learning environments. While the review identifies significant technical advances, it remains focused on design-level opportunities. It does not address how such architectures could be translated into operational auditing frameworks involving external auditors or regulators. Moreover, questions related to the verification of logs, real-world scalability, and privacy trade-offs remain relatively unexplored.

By analysing these proposals, several gaps persist. Most notably, few systems are designed with multistakeholder governance in mind. While blockchain can decentralise control, most implementations still centre the AI provider as the sole writer and gatekeeper of audit records. Additionally, post-deployment transparency, such as logging inference outcomes or downstream decisions, is rarely supported, despite being essential for real-time accountability, mainly in high-risk domains. Finally, there is limited alignment with legal auditability standards, such as those required by the EU AI Act, which mandates risk classification, conformity assessments, and transparency measures for both developers and deployers.

The framework proposed in this paper addresses these limitations by (i) embedding cryptographic verifiability and provenance tracking across the AI lifecycle, (ii) supporting selective transparency and access control for legal and ethical compliance, and (iii) enabling distributed audit governance where third parties, such as regulators, civil society, and academic and technical experts can independently verify claims. This moves beyond conceptual logging toward institutionalised, legally compatible, and socially accountable AI auditability.

3 Conceptual Framework: Blockchain-Mediated AI Audit

3.1 Principles

Our framework proposes four central principles that address the observed limitations in current auditing mechanisms.

The first principle advocates for comprehensive lifecycle coverage, in contrast to the fragmented nature of current auditing approaches. Unlike existing mechanisms, which focus on development or deployment phases in isolation, our framework captures audit artefacts throughout the entire AI lifecycle (Brundage, Avin, Wang, Belfield, Krueger, Hadfield et al., 2020). This perspective begins with data provenance records that document the origin, licensing terms, and pre-processing steps applied to training data, creating an immutable record of the fundamental inputs that define model behaviour.

Subsequently, model training artefacts must be recorded, which will serve as verifiable evidence of development steps, deployment, and behaviour during interaction with users. Various relevant data should be logged, including:

- Checkpoints: Hashes of AI model weights and hyperparameters, along with validation metrics and version control information, enabling technical scrutiny and independent reproducibility.
- Inference Events: Detailed records of AI model predictions or decisions, with an emphasis on high-risk cases such as medical diagnostics and legal decisions.
- Model Updates: Comprehensive logs of changes made to the model, including their justification, new performance metrics, and the identification of parties responsible for proposing or approving these changes.
- Data Provenance Records: A detailed history of all data used in model development, as well as the identification of the origin and nature of training datasets.
- Interactions: Records of user interactions on sensitive topics, including feedback annotated on the collaborative platform by public and private reviewers.
- Compliance Events: Hashes of compliance evaluations and risk classifications, for verifying the model's adherence to specific regulatory requirements.

The second principle is that of selective transparency with privacy preservation, which acknowledges that not all audit information should be publicly accessible. This principle provides a balanced scenario between two competing precepts: transparency and confidentiality (Koshiyama, Kazim, and Treleaven, 2021). Our framework implements a layered access model, allowing appropriate visibility for different actors according to pre-established governance rules.

The public layer contains aggregated performance metrics, bias indicators, and high-level model cards, without exposing confidential details. The regulatory layer grants authorities access to detailed logs and training data statistics, information necessary for verifying their respective origins and potential compliance investigations. Finally, the provider layer encompasses proprietary implementation details and all raw data used in training, which are accessible only to their owners, unless specific legal provisions require some form of disclosure.

This layered approach is implemented through a combination of on-chain smart contract access control and off-chain encrypted storage, with zero-knowledge proofs enabling the verification of claims without requiring full disclosure of the underlying data (Benarroch, Campanelli, Fiore et al., 2023).

The third principle corresponds to cryptographic verifiability, ensuring that all audit claims can be independently verified. Thus, each audit artefact is linked to cryptographic hashes, creating an immutable chain of evidence that prevents tampering or post-hoc revisions (Xu, Weber, and Staples, 2019).

The established standard uses SHA-256 hashes, which create unique digital fingerprints capable of detecting any subsequent modification. Digital signatures from authorised parties, such as developers, auditors, and regulators, provide a non-repudiation condition, guaranteeing the authorship of actions.

The use of Merkle trees allows for batch verification of all artefacts, which are cryptographically dependent on each other. Automatic timestamps establish a definitive and immutable temporal order, which is essential for determining causality and attesting to eventual compliance with regulatory requirements.

The fourth and final principle is that of multistakeholder governance, which recognises the inherent power asymmetries in provider-controlled auditing systems. The proposed framework incorporates governance mechanisms directly embedded within smart contracts, ensuring that rules are automatically

enforced without the need for any voluntary manifestation (Filippi and Wright, 2018), embracing the logic of "Code is Law" (Lessig, 1999).

However, we acknowledge the critiques of technological solutionism inherent in this phrase. Our framework interprets "Code is Law" not as a replacement for legal oversight, but as a tool to enforce predefined governance rules transparently. This on-chain enforcement must be complemented by robust off-chain legal processes to address power imbalances, ensure equitable participation, and allow for the evolution of rules, a concept aligned with the principle of regulatory equivalence, where technological systems and legal frameworks mutually reinforce one another.

Beyond technical decentralisation, involving multiple stakeholders enhances the perceived legitimacy of AI auditing processes and strengthens public trust. By embedding governance rules in smart contracts and granting oversight powers to actors beyond AI providers, the framework aligns with established principles of procedural fairness and participatory regulation.

A predefined and well-established set of permissions determines which actors can read and write audit records, preventing unauthorised access and any attempt at unilateral censorship. The creation of integrated dispute resolution mechanisms allows stakeholders to submit formal challenges against AI model decisions, triggering automatic and transparent investigation processes that cannot be suppressed.

Finally, voting protocols can be developed to facilitate the updating of audit standards as technological and regulatory requirements evolve, ensuring the long-term relevance of the proposed framework.

3.2 On-chain vs. Off-chain Architecture

The inherent decentralisation of blockchain technology is counterbalanced by its inefficiency in large-scale scenarios, as logging all information related to AI model auditing is impractical due to the involved costs and high latency in the case of complex queries.

Aware of these limitations, the proposed framework utilises a hybrid model, wherein critical and sensitive metadata, along with their respective cryptographic proofs, are stored on-chain. In contrast, large volumes of information, such as training datasets, their original weights, detailed inference logs, and all AI model documentation, will be stored off-chain in traditional, scalable, and lower-cost solutions, such as IPFS (InterPlanetary File System) or secure commercial databases.

On-chain storage is designated for data requiring the highest levels of integrity and availability, such as hashes of audited artefacts, digital signatures of responsible parties, and access control lists, ensuring the impossibility of fraud or any attempt to alter records.

In compliance with privacy regulations (e.g., GDPR), all personal data must be encrypted before off-chain storage. The proposed hybrid approach meets regulatory privacy standards and maintains the integrity of the audit trail through rewritten on-chain records (Politou, Alepis, and Patsakis, 2018). The immutable nature of blockchain can create tensions with GDPR rights such as the right to erasure (Art. 17). Our hybrid design mitigates this by keeping personal data off-chain and using on-chain pointers and hashes. When data deletion is required, the off-chain content can be effectively erased, and the corresponding on-chain hash rendered unusable, achieving functional compliance while preserving the integrity of the audit trail.

3.3 Threat Models

All proposed technological frameworks are subject to various threats, both from malicious internal actors and third parties. AI models are particularly affected by data poisoning attempts aimed at introducing biases into interactions or corrupting their accuracy, as well as techniques like jailbreaking, which are particularly critical for our proposal, as they aim to circumvent ethical guidelines or obtain responses to illicit topics. Other threats, such as adversarial machine learning or model extraction attacks via API queries, must also be considered.

While blockchain guarantees the integrity of audit records once they are logged, it does not automatically ensure the quality of the underlying data or audits. To address the so-called 'garbage in, garbage out' problem, the framework embeds upstream verification layers: datasets and audit inputs are subject to provenance checks and cryptographic attestations by multiple independent stakeholders before being anchored on-chain. This reduces the likelihood of corrupted or poor-quality audits becoming immutable records, allowing subsequent actors to trace responsibility to the original sources.

Blockchain technology is also subject to considerable risks that can corrupt the integrity of audit logs by exploiting potential vulnerabilities in smart contracts or attempts to manipulate consensus protocols that ensure data logging and block finalisation.

Our security model assumes that the majority of blockchain nodes are honest actors, such that at least 51% operate following the specifications of the consensus protocol (Garay, Kiayias, and Leonardos, 2015). In permissioned environments such as Hyperledger Fabric, collusion between participating nodes is a realistic threat. To mitigate this, node selection should privilege institutions with distinct mandates and accountability mechanisms (e.g., regulators, ethics boards, and academic auditors), thus reducing the likelihood of coordinated malicious behaviour. Additional safeguards include periodic external audits of the node governance process and dynamic membership policies that enable the removal of compromised nodes.

4 Evaluating the proposed framework: pathways to empirical testing and pilots

To move from conceptual design to practical implementation, the proposed blockchain-based, multistakeholder AI auditing framework must be tested for effectiveness across several dimensions: transparency, accountability, regulatory compliance, scalability, and stakeholder usability. This section outlines how the framework can be empirically evaluated in both controlled settings and real-world pilot deployments.

The EU AI Act mandates enhanced auditability for high-risk AI systems. These domains provide ideal testbeds for the proposed framework, particularly in applications where algorithmic decisions affect access to resources or fundamental rights. A pilot study could, for instance, involve deploying the audit infrastructure within a hospital's diagnostic AI pipeline, logging model version histories, data input records, and inference outputs immutably on-chain. Performance could be assessed on metrics such as audit completeness, retrieval latency, and external verifiability.

Comparable small-scale pilots have already demonstrated feasibility. Martinez, Magdalena, and Savitri (2024) used a blockchain layer to audit AI-generated financial transactions, showing gains in both traceability and fraud prevention.

Our framework builds upon these efforts by integrating regulatory obligations and a governance model, which would also need to be evaluated.

In the governance field, another key evaluation method involves simulating the multistakeholder governance process under different operational conditions. Role-play exercises or agent-based models can be used to test whether civil society organisations, regulators, AI providers, and auditors can effectively interact using the proposed access-control policies and smart contract–based dispute resolution mechanisms. Key questions would include: Do stakeholders understand and trust the evidence logged? Can they effectively contest or validate claims? Does the system distribute audit authority equitably?

Regarding usability, it is a critical and often overlooked aspect of blockchain-based governance systems. The proposed framework should undergo formal user testing involving auditors, compliance officers, and potentially affected users. For instance, participants could be asked to trace an AI model's decision history or verify the authenticity of an audit log. Success would be measured through usability metrics such as task completion time, error rates, and subjective trustworthiness (e.g., via Likert-scale feedback).

Finally, regulatory alignment could be tested through simulated compliance audits based on the EU AI Act's Article 50 (transparency obligations) and Article 72 (post-market monitoring). Legal and technical experts could assess whether the logs generated by the system meet statutory requirements for traceability, explainability, and data protection.

In sum, empirical validation of the framework should take a multi-pronged approach: sector-specific pilots, stakeholder governance simulations, usability studies, and regulatory audit simulations. Together, these methods will provide a robust understanding of how the framework performs as a technical infrastructure and as a legally compatible and socially inclusive audit mechanism.

5 Conclusion, Limitations & Future Work

As AI systems become increasingly embedded in high-stakes decision-making domains, the demand for robust, transparent, and accountable audit infrastructures has grown correspondingly urgent. Existing audit mechanisms represent important initial steps but fall short in offering persistent, verifiable, and democratically governed oversight. They often rely on centralised, opaque processes, lack enforcement

mechanisms, and exclude meaningful participation by external stakeholders, particularly those most impacted by AI decisions.

This article has argued that blockchain technology, when appropriately tailored to the needs of AI governance, offers a robust suite of capabilities to address these gaps. By providing immutability, decentralisation, and cryptographic verifiability, blockchain can anchor audit records in tamper-proof infrastructures, ensure pluralistic control over audit processes, and facilitate compliance with evolving regulatory frameworks like the EU AI Act, OECD Principles, and others. Crucially, the proposed framework advances beyond prior work by integrating multistakeholder governance, lifecycle-wide traceability, and selective transparency via smart contract–based access control.

The outlined framework addresses both technical and institutional shortcomings identified in the literature. It offers a hybrid on-chain/off-chain audit design that enables secure logging of key development and deployment events, supports legally relevant provenance tracking, and provides verifiable decision logs for real-time or retrospective auditing. A pilot-ready evaluation strategy has also been proposed, including simulation environments, sector-specific trials, and usability assessments across diverse stakeholder groups.

Moreover, the framework directly addresses the information asymmetry highlighted in Section 1 by shifting audit control from unilateral AI providers to distributed, verifiable ledgers. Rather than relying on providers' voluntary disclosures, the system makes transparency a structural property of the infrastructure itself, thereby closing the gap between identified audit failures and proposed solutions.

Finally, while the paper focuses primarily on auditability, the same blockchain infrastructure could support traceability of multilingual datasets and the inclusion of underrepresented languages. By recording provenance information for minority-language data contributions, blockchain can help ensure their visibility and accountability within AI training pipelines, addressing the linguistic asymmetries raised in the Introduction.

While promising, the framework's real-world implementation will face challenges in interoperability, standard-setting, and balancing audit transparency with data protection laws like the GDPR. These tensions present areas for future research and policy collaboration.

Ultimately, this work contributes toward transforming AI audits from a fragmented, black-box process into an open, verifiable, and participatory system. By fusing the institutional ambitions of emerging AI regulation with the technical guarantees of blockchain, it lays the groundwork for a more trustworthy and democratically accountable AI ecosystem.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ada Lovelace Institute (2022) Algorithmic impact assessment: User guide. Tech. rep., Ada Lovelace Institute, URL https://www.adalovelaceinstitute.org/resource/aia-user-guide/

Ahmed S (2025) Enhancing data security and transparency: The role of blockchain in decentralized systems. International Journal of Advanced Engineering, Management and Science 11(1):167–176. 10.22161/ijaems.111.12

Akther A, Arobee A, Adnan AA, et al (2025) Blockchain as a platform for artificial intelligence (AI) transparency. Version 1, arXiv:2503.08699

Aldeco-Pérez R, Moreau L (2010) Securing provenance-based audits. In: McGuinness DL, Michaelis JR, Moreau L (eds) Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science, vol 6378. Springer Berlin Heidelberg, Berlin, Heidelberg, p 148–164, 10.1007/978-3-642-17819-1_18

Balan K, Learney R, Wood T (2025) A framework for cryptographic verifiability of end-to-end AI pipelines. In: Proceedings of the 10th ACM International Workshop on Security and Privacy

Analytics, pp 49–59, 10.1145/3716815.3729011

Benarroch D, Campanelli M, Fiore D, et al (2023) Zero-knowledge proofs for set membership: efficient, succinct, modular. Designs, Codes and Cryptography 91:3457–3525. 10.1007/s10623-023-01245-1

Bhat A, Coursey A, Hu G, et al (2023) Aspirations and practice of ml model documentation: Moving the needle with nudging and traceability. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '23, 10.1145/3544548.3581518, URL https://doi.org/10.1145/3544548.3581518

Birhane A, Steed R, Ojewale V, et al (2024) AI auditing: The broken bus on the road to AI accountability. In: 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pp 612–643, 10.1109/SaTML59370.2024.00037

Bracamonte V, Pape S, Löbner S, et al (2023) Effectiveness and information quality perception of an AI model card: A study among non-experts. In: 2023 20th Annual International Conference on Privacy, Security and Trust (PST), pp 1–7, 10.1109/PST58708.2023.10320197

Brundage M, Avin S, Wang J, et al (2020) Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv:2004.07213

Busuioc M, Curtin D, Almada M (2023) Reclaiming transparency: Contesting the logics of secrecy within the AI act. European Law Open 2(1):79–105. 10.1017/elo.2022.47

Butt A, Junejo AZ, Ghulamani S, et al (2023) Deploying blockchains to simplify AI algorithm auditing. In: 2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp 1–6, 10.1109/ICETAS59148.2023.10346420

Caparroz R (2024) A necessidade de criação de modelos de inteligência artificial voltados à realidade brasileira. In: Inteligência Artificial: Desafios para a realidade brasileira. IBChain, São Paulo

Casper S, Ezell C, Siegmann C, et al (2024) Black-box access is insufficient for rigorous AI audits. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp 2254–2272, 10.1145/3630106.3659037

Filippi PD, Wright A (2018) Blockchain and the Law: The Rule of Code. Harvard University Press, Cambridge, MA, URL http://www.jstor.org/stable/j.ctv2867sp

Garay J, Kiayias A, Leonardos N (2015) The bitcoin backbone protocol: Analysis and applications. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Berlin, Heidelberg, pp 281–310

Habibikilak F (2024) Examination of the role of ethics-based auditing in the development of trustworthy artificial intelligence (case study: Auditing firms in tehran). International Journal of Innovation Management and Organizational Behavior 4(4):231–238. 10.61838/kman.ijimob.4.4.27

International Association of Privacy Professionals (IAPP) (2025) Artificial Intelligence Governance Professional. URL https://iapp.org/certify/aigp/, acessado em: 09 de outubro de 2025

ISACA (2025) Artificial Intelligence Auditing Associate. URL https://www.isaca.org/credentialing/aaia, acessado em: 09 de outubro de 2025

ISO (2023) ISO/IEC 42001:2023. Tech. rep., International Organization for Standardization, URL https://www.iso.org/standard/81230.html

Kalapaaking AP, Khalil I, Yi X, et al (2025) Auditable and verifiable federated learning based on blockchain-enabled decentralization. IEEE Transactions on Neural Networks and Learning Systems 36(1):102–115. 10.1109/tnnls.2024.3407670

Kempton A, Parmiggiani E, Vassilakopoulou P (2023) Accountability in managing artificial intelligence: State of the art and a way forward for information systems research. In: ECIS 2023 Research Papers, URL https://aisel.aisnet.org/ecis2023rp/361/, paper 361

Kiseleva A, Kotzinos D, De Hert P (2022) Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations. Frontiers in Artificial Intelligence 5:879603. 10.3389/frai.2022.879603

Koshiyama A, Kazim E, Treleaven P (2021) Algorithm auditing: Managing the legal, ethical and technological risks of artificial intelligence, machine learning and associated algorithms. Computer 55(4):40–50. 10.1109/MC.2021.3067225

Lam K, Lange B, Blili-Hamelin B, et al (2024) A framework for assurance audits of algorithmic systems. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp 1078–1092, 10.1145/3630106.3658957

Lessig L (1999) Code and Other Laws of Cyberspace. Basic Books, New York, NY

Liao QV, Subramonyam H, Wang J, et al (2023) Designerly understanding: Information needs for model transparency to support design ideation for AI-powered user experience. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp 1–21, 10.1145/3544548.3580652

Martinez D, Magdalena L, Savitri AN (2024) AI and blockchain integration: Enhancing security and transparency in financial transactions. International Transactions on Artificial Intelligence 3(1):11–20. 10.33050/italic.v3i1.651

Mitchell M, Wu S, Zaldivar A, et al (2019) Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp 220–229, 10.1145/3287560.3287596

Nakamoto S (2008) Bitcoin: A peer-to-peer electronic cash system. URL https://bitcoin.org/bitcoin.pdf

Ojewale V, Steed R, Vecchione B, et al (2025) Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp 1–29, 10.1145/3706598.3713301

Pasquale F (2016) The Black Box Society: The Secret Algorithms that Control Money and Information. Harvard University Press, Cambridge, MA

Pérez B, Rubio J, Sáenz-Adán C (2018) A systematic review of provenance systems. Knowledge and Information Systems 57(3):495–543. 10.1007/s10115-018-1164-3

Politou E, Alepis E, Patsakis C (2018) Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. Journal of Cybersecurity 4(1):tyy001. 10.1093/cybsec/tyy001

Potdukhe N, Gourshettiwar P, Zade S, et al (2025) Blockchain-enabled federated learning systems with explainable AI: A review. In: 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS), pp 1650–1656, 10.1109/icmlas64557.2025.10968523

Raji ID, Xu P, Honigsberg C, et al (2022) Outsider oversight: Designing a third party audit ecosystem for AI governance. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp 557–571, 10.1145/3514094.3534181

Ramos S, Ellul J (2024) Blockchain for artificial intelligence (AI): Enhancing compliance with the EU AI act through distributed ledger technology. a cybersecurity perspective. International Cybersecurity Law Review 5(1):1–20. 10.1365/s43439-023-00107-9

Reisman D, Schultz J, Crawford K, et al (2018) ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY. Tech. rep., AI Now

Institute, URL https://ainowinstitute.org/wp-content/uploads/2023/04/aiareport2018.pdf

Severino F, Canciani A, Felicioli C, et al (2024) Trustworthy AI for infrastructure monitoring: A blockchain-based approach. Procedia Structural Integrity 62:276–284. 10.1016/j.prostr.2024.09.043

Shen H, Wang L, Deng WH, et al (2022) The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In: 2022 ACM Conference on Fairness Accountability and Transparency, pp 440–451, 10.1145/3531146.3533110

Sutton A, Samavi R (2018) Tamper-proof privacy auditing for artificial intelligence systems. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp 5374–5378, 10.24963/ijcai.2018/756

Targino N, Souza D, Salgado AC (2024) An approach to associate provenance from ML pipelines with transparency and accountability. In: 2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA), pp 1–8, 10.1109/IISA62523.2024.10786622

Treasury Board of Canada Secretariat (2024) Algorithmic impact assessment tool. URL https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Urman A, Smirnov I, Lasser J (2024) The right to audit and power asymmetries in algorithm auditing. EPJ Data Science 13(1):19. 10.1140/epjds/s13688-024-00454-5

Williams R, Cloete R, Cobbe J, et al (2022) From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy 4:e7. 10.1017/dap.2021.37

Xu X, Weber I, Staples M (2019) Architecture for blockchain applications. Springer, Cham, Switzerland

Yadav AS, Singh N, Kushwaha DS (2023) Evolution of blockchain and consensus mechanisms & its real-world applications. Multimedia Tools and Applications 82(22):34363–34408. 10.1007/s11042-023-14624-6

Zheng Z, Xie S, Dai H, et al (2017) An overview of blockchain technology: Architecture, consensus, and future trends. In: 2017 IEEE International Congress on Big Data (BigData Congress), pp 557–564, 10.1109/BigDataCongress.2017.85

Bootstrapping Trust across Web2 and Web3 Domains Using Publicly Verifiable Web Data

Yuan Lu[1], Qiang Tang[2]

[1] Institute of Software, Chinese Academy of Sciences, Beijing, China.
[2] School of Computer Science, University of Sydney, Australia.

Corresponding authors. E-mails: luyuan@iscas.ac.cn; qiang.tang@sydney.edu.au

Abstract

Through public key infrastructure (PKI) and the TLS protocol, users can securely connect to trusted web services such as banks, e-commerce platforms, and government portals. If such data could be made publicly verifiable, it could serve as a natural trust anchor for decentralized Web3 systems. Prior efforts, such as DECO (Zhang et al., CCS'20), introduced TLS web oracles that allow a prover to convince a designated verifier that specific data was obtained from a TLS server. However, these designs rely on the verifier's cooperation and allow strategic aborts that prevent the prover from demonstrating the authenticity of already-transmitted TLS data.

We present a new TLS oracle protocol that achieves public verifiability and a novel form of fairness using an "encrypt-TLS-keys-to-future" paradigm. The prover and verifier jointly emulate a TLS client via secure two-party computation (2PC), obtaining additive shares of the session key. They also set up a trapdoor verifiable delay function (tVDF), allowing them to encrypt the full TLS session key under a future-unlockable value, while the TLS ciphertext is recorded on-chain. After a sufficiently long delay, the prover can reconstruct the session key and use zero-knowledge proofs to selectively disclose committed TLS data.

As an application, we present the first decentralized protocol that enables fair exchange between blockchain-backed Web3 assets and TLS-protected Web2 assets, bridging the gap between the two domains by supporting asset swaps without relying on any trusted third-party mediators.

Keywords

TLS Oracle, Fair Exchange, P2P Swap, Web2-Web3 Interoperability.

1 Introduction

Decentralized systems promise a "permissionless" world where individuals interact, transact, and collaborate without centralized supernodes. This vision fosters the Web3 movement, where centralized applications are built atop blockchain and operate under rules enforced by distributed consensus (Labs, 2017; Nakamoto, 2008; Wood et al., 2014). However, such a trend of decentralization introduces a new problem: the absence of native trust anchors. In particular, Web3 participants typically interact using cryptographic pseudonyms, which are long-lived public keys derived from wallet addresses or smart contract identities. While this design choice enhances privacy and censorship resistance, it also presents significant challenges. Many traditional forms of trade and transactions that are straightforward in Web2 (due to real-world identities, reputation systems, and legal frameworks) become much more challenging to realize in the permissionless environment of Web3, due to the lack of mechanisms linking malicious behaviors to real-world identities.

Nevertheless, the current Internet (commonly referred to as Web2) has established trust through its well-established infrastructure. For example, when browsing HTTPS-enabled web pages (Rescorla, 2000), the browser displays a lock icon to indicate that the user is securely connected to a verified service. This is enabled by the TLS protocol (Dierks and Rescorla, 2008; Rescorla, 2018), which is backed by digital certificates issued by certificate authorities (CAs) as part of the broader Public Key Infrastructure (PKI) (Cooper, Santesson, Farrell, Boeyen, Housley, and Polk, 2008; ISO, 2022). For example, when users log into digital banking platforms, government portals, or shopping websites, they rely on HTTPS running over TLS to authenticate the server and ensure data integrity and confidentiality; Moreover, many of these web services will authenticate users in return via passwords and other second-factor authenticators, after which they can display pages containing versatile identity-linked data that serve as reputation profile information, such as names, dates of birthdays, social security numbers, account balances, transaction histories, credit scores, and more.

The above discrepancy between Web2 and Web3 raises a natural question:

Can we leverage the existing trust infrastructure of the traditional Internet to bootstrap cross-domain trust in the decentralized systems? Specifically, can we ensure that web users extract verifiable claims from trusted Internet services and anchor them on blockchain platforms in a way that is privacy-preserving and publicly verifiable?

Prior Studies and Their Limitations. At first glance, it may seem straightforward to modify the server-side TLS protocol to support publicly verifiable extraction of TLS data. For example, a website could directly sign user-specific data, enabling smart contracts to verify authenticity without intermediaries (Backman, Richer, and Sporny, 2024). However, this approach is largely impractical, since most Web2 services, especially large organizations such as governments, banks, and major e-commerce platforms, have little incentive to alter their infrastructure to accommodate the demand of Web3 world. Such modifications could introduce compatibility challenges, operational costs, and regulatory risks, while offering no direct benefit or revenue to their existing business models.

Several standard-compatible efforts are also made to bridge Web2's trust infrastructure and Web3, proposing TLS web oracle systems, which extract authenticated web data for verification by a designated verifier (typically referred to as an oracle service provider). A pioneering example is Town Crier (Zhang, Cecchetti, Croman, Juels, and Shi, 2016), which leverages Intel SGX to run a trusted enclave that fetches HTTPS-protected data and relays the result accompanied by SGX remote attestation to smart contracts. The security of Town Crier critically depends on the non-breakability of SGX, assuming that the enclave operates correctly and remains secure against hardware-level attacks. However, this trust model could be elusive in practice: First, SGX and similar trusted execution environments (TEEs) have a growing list of known vulnerabilities, which undermines their long-term reliability; Second, the model implicitly requires trust in Intel, the centralized hardware vendor, for correct remote attestation. These concerns highlight the fragility of hardware-rooted trust in decentralized settings.

To eliminate the reliance on trusted hardware, DECO (Zhang, Maram, Malvai, Goldfeder, and Juels, 2020) introduced a cryptographic approach that leverages zero-knowledge proofs over standard TLS sessions. In DECO, a prover engages in a TLS connection with a web server (and a designated verifier), thus finally producing a zero-knowledge proof attesting that specific data was received from the authenticated server. While DECO eliminates the need for trusted hardware such as SGX, it still depends on a

designated verifier who has the exclusive capacity to validate the proof. If the goal is to interface with a public blockchain, the designated verifier finally needs to publish the authenticated data to the public. This publishing step is not publicly verifiable and reintroduces a centralized point of failure. Further improvements have emerged. For example, Janus (Lauinger, Ernstberger, Finkenzeller, and Steinhorst, 2025) offers privacy-preserving provenance for the latest version of TLS 1.3 with special optimizations to the cipher suites of TLS 1.3. Xie et al. propose the garble-then-prove framework (Xie, Yang, Wang, and Yu, 2024) to avoid malicious-secure garbling circuits. However, these approaches mostly follow the design framework from DECO, and therefore, they are still designed for a designated verifier and do not support public verifiability in the presence of a dishonest verifier.

While in principle, the aforementioned designs (Lauinger et al., 2025; Xie et al., 2024; Zhang et al., 2020) can be extended to support on-chain verifiability, for example, by having the prover commit ciphertexts to the blockchain and later reconstruct the TLS session key with the verifier's help. However, such adaptations remain fundamentally limited. In particular, they cannot guarantee that the prover can generate a valid proof for already-transmitted data if the verifier chooses to abort or becomes unresponsive after the transmission. This shortcoming is especially problematic in applications such as cross-domain exchange, where one party (e.g., the verifier) may have already received off-chain value (e.g., fiat payment) but can strategically withhold their key share to prevent the other party from completing a proof for the successful completion of off-chain transactions. In such cases, the lack of fairness in enabling verifiability opens the door to asymmetric outcomes and undermines mutual trust in such decentralized systems.

In summary, despite meaningful progress, existing designs (Backman et al., 2024; Lauinger et al., 2025; Xie et al., 2024; Zhang et al., 2016,2) fall short of meeting the core requirements of Web3: (1) compatibility with standard TLS servers, (2) avoidance of strong or impractical trust assumptions, and (3) support for proof and public verifiability of any already-transmitted TLS data. Our work aims at addressing these limitations by introducing a novel "encrypt-TLS-key-to-future" paradigm, in which jointly emulated TLS clients defer TLS key recovery at the prover side through a trapdoor verifiable delay function (tVDF), and later produce publicly verifiable zk-SNARKs attesting to the authenticity of the extracted web data.

2 Results

This paper addresses these limitations of previous Web oracle protocols (Backman et al., 2024; Lauinger et al., 2025; Xie et al., 2024; Zhang et al., 2016,2), by proposing a new approach that avoids reliance on trusted third parties for public access to the extracted authenticated web data. In particular, our contribution is three-fold:

- At the core of our design, we introduce a new technique paradigm called "encrypt-TLS-keys-to-future" for realizing a TLS web oracle. In this approach, a prover and a verifier collaboratively emulate a TLS client using secure two-party computation (2PC), then establish a connection with a TLS server to obtain additive shares of the resulting session keys, which is so far similar to prior designs such as DECO. The key innovation of the "encrypt-TLS-keys-to-future" paradigm lies in the following: the verifier and the prover distributedly set up a trapdoor verifiable delay function (tVDF) (Wesolowski, 2020) and distributedly encrypt the TLS session keys using the tVDF output obtained from the trapdoor evaluation of tVDF. After a predetermined and sufficiently large time delay, the prover can eventually unlock the tVDF to decrypt the TLS session keys. This enables the prover to be able to *solely* decrypt the ciphertext of TLS transcripts that are already committed to the blockchain.
- Another contribution is a new oracle protocol design, which is publicly verifiable and supports selective disclosure, based on the "encrypt-TLS-keys-to-future" paradigm.[91] An observation from our "encrypt-TLS-keys-to-future" paradigm is: the prover eventually can decrypt the TLS ciphertext committed to the blockchain, which means the prover can leverage generic zero-knowledge proof scheme, using the TLS session key as witness, to generate a zero-knowledge proof attesting selective disclosure of specific data fields within the encrypted transcript (in lieu of decrypting the whole

[91] Note that there may be alternative ways to modify existing TLS oracle protocols to support public on-chain verification. However, our design from the "encrypt-TLS-keys-to-future" paradigm offers a unique feature: if certain events such as fiat money transfers occur during the TLS session, the prover alone can later prove that these transmissions took place, even if the verifier takes no further action. This property is particularly important in fair exchange settings, where relying on the verifier's continued participation may lead to unfairness (since the proof for TLS data extraction needs the verifier's local secret to compute). We will present a concrete example demonstrating the usefulness of this property in Section 5.

ciphertext), thereby publicly demonstrating the authenticity of certain data retrieved during the TLS session without revealing anything else.

A unique feature of the resulting TLS oracle protocol is a form of "fairness": if some event indeed occurs within the TLS session (e.g., a certain web data was transmitted), the prover alone can generate a proof to enable the public blockchain to verify the authenticated data from the TLS session, even if the other verifier maliciously aborts or becomes unresponsive.

- Finally, to show the power of our TLS oracle design from the "encrypt-TLS-keys-to-future" paradigm, we demonstrate how this publicly verifiable oracle can be applied to enable decentralized fair exchange between blockchain-backed cryptocurrency (e.g., Ethereum tokens (Wood et al., 2014)) and TLS-backed digital assets (e.g., bank balances) on the Internet. To our knowledge, this is the first decentralized protocol enabling the fair exchange between off-chain and on-chain assets in the absence of a trusted third-party mediator (Chen, Liu, Lu, and Tang, 2025).

  This instantiation for on-chain and off-chain exchange showcases the viability of our publicly verifiable web oracle design in bridging the existing Web2 infrastructure to the decentralized Web3 world, in a secure and decentralized manner.

3 Cryptographic Primitives

In this section, we will briefly recall the key cryptographic components that are relevant to our public-verifiable Web oracle design.

3.1 Digital Certificate and Public Key Infrastructure

Public Key Infrastructure (PKI) (Cooper et al., 2008; ISO, 2022) is the foundational trust system that underpins secure communication across the Internet, particularly used throughout the Transport Layer Security (TLS) protocol to verify the authenticity of Web servers. Specifically, PKI enables the user clients to authenticate web servers using digital certificates that bind a domain name to a public key. This mechanism is essential for ensuring that users are communicating with the intended server rather than a phishing website. A digital certificate (Cooper et al., 2008) of some domain name (e.g., www.bankofamerica.com) is issued by a Certificate Authority (CA), which is a trusted entity verifying the domain owner's identity and subsequently issues the certificate containing, but not limited to, the following information:

- The server's public key;
- The server's identity (e.g., domain name);
- The expiration time of the certificate;
- The CA's identity and its digital signature over the certificate's contents.

When a client connects to a server via TLS, the server presents its certificate during the handshake phase. Therefore, the client can authenticate the server by verifying that the server's certificate was issued by a CA it trusts (e.g., signed by a root CA hardcoded in the client's operating system or browser, or by an intermediate CA that links back to a root CA through a valid certificate chain).

3.2 Transport Layer Security (TLS)

Given trust over the already established PKI system, clients can further leverage the TLS protocol to connect to their intended website (though only the clients themselves can verify the website's authenticity, and our goal is to extend such designated clients' verifiability to public verifiability). For a better understanding of the context, we briefly recall the workflow and components of TLS.

From a high level, TLS operates in two phases (Dierks and Rescorla, 2008; Rescorla, 2018): handshake and data transmission. In the handshake, the client and server perform an authenticated key exchange to establish a shared pre-master secret, which is used to derive the (symmetric) session keys for the subsequent data transfer. Then, application data is transmitted securely with confidentiality and integrity, using the session keys. Notably, the server-to-client key (i.e., server's key) and the client-to-server key (i.e., client's key) can be thought of being independently derived from the pre-master secret.

Note that TLS has various versions and offers multiple choices of cipher suits. Here in this work, we will instantiate our design for TLS 1.2 (Dierks and Rescorla, 2008; Nir, Josefsson, and Pégourié-Gonnard, 2018; Salowey, Choudhury, and McGrew, 2008), using (i) elliptic curve DH key exchange (ECDHE) for

establishing the pre-master secret (Nir et al., 2018), (ii) SHA256-based message authentication code (HMAC) for deriving session keys from the pre-master secret (Krawczyk, Bellare, and Canetti, 1997), and (iii) AES-GCM or other authenticated encryption schemes for secure transfer of payload data (Dworkin, 2007; Dworkin, Barker, Nechvatal, Foti, Bassham, Roback, Dray Jr et al., 2001; McGrew, 2008; Rogaway, 2002). At the time of writing, TLS 1.2 with ECDHE, AES128-GCM, and SHA256 remains one of the most widely supported TLS cipher suites in real-world deployments (Qualys, ????). Moreover, our implementation can be seamlessly extended to support TLS 1.3 (Rescorla, 2018) with AES256-GCM and SHA384 without fundamental barriers.

Below, we briefly review these TLS components.

Negotiating shared secret from ECDHE. In ECDHE (Nir et al., 2018), the client chooses a random $c \leftarrow_{\$} \mathbb{F}_q$ and sends $g^c \in \mathbb{G}_q$ to the server, and the server chooses a random $s \leftarrow_{\$} \mathbb{F}_q$ and sends $g^s \in \mathbb{G}_q$ to the client, such that both of them can compute the common group element $Y = g^{cs} \in \mathbb{G}_q$. Here $\mathbb{F}_q$ is a prime field and $\mathbb{G}_q$ is a cyclic group with order q. The first coordinate of Y, $Y[0]$, is then used as the pre-master secret. In TLS, ECDHE is also enhanced to be authenticated. In our setting, the server authenticates itself by showing its digital certificate and a digital signature.

Key derivation from HMAC. Both TLS 1.2 and 1.3 implement their key derivation processes based on HMAC (Krawczyk et al., 1997), with differences in concrete instantiations. In particular, to derive symmetric session keys, TLS 1.2 realizes a hash-based pseudo-random function (HPRF) from HMAC; it then uses the pre-master secret $Y[0]$ as the HPRF key, on input the sum of client's nonce and server's nonce with a label "master secret", to compute the so-called master secret ms; after that, it uses HPRF again, keyed by ms with a label "key expansion", to derive the symmetric session keys.

Authenticated encryption from AES-GCM. AES is a standardized block cipher scheme (Dworkin et al., 2001), and its Galois/Counter mode (GCM) (Dworkin, 2007) extends it to provide authenticated encryption with associated data (AEAD) (McGrew, 2008; Rogaway, 2002). Informally speaking, AEAD not only encrypts messages to ensure their confidentiality against an adversary not knowing the encrypting key, but also generates the tags of message authentication code (MAC) to ensure the integrity of messages, thus preventing an adversary from tampering messages without access to the AES-GCM key.

3.3 Trapdoor Verifiable Delay Function

Our protocol employs an advanced cryptographic primitive known as a verifiable delay function (VDF) (Boneh, Bonneau, Bünz, and Fisch, 2018; Pietrzak, 2019; Wesolowski, 2020), equipped with trapdoor evaluation (Wesolowski, 2020) (tVDF), to realize the "encrypt-TLS-keys-to-future" paradigm. At a high level, a VDF is a function whose output must take a minimum amount of sequential time to compute, no matter how much parallel computing power is applied and what arbitrary acceleration algorithms are adopted. However, once the output is computed, it is accompanied by a proof for correctness, such that anyone (including smart contracts) can verify its correctness efficiently. A trapdoor VDF (Wesolowski, 2020) further includes a trapdoor that allows its holder to bypass the sequential delay and compute the output almost instantly. Later in our protocol, we leverage this trapdoor to encrypt the TLS session keys in advance, before anyone (without the trapdoor) can feasibly compute the VDF output through the standard slow path. This ensures that the decryption of TLS traffic can happen and only happen after a designated delay, enforcing a form of time-lock and enabling delayed verifiability.

3.4 Zero-Knowledge Succinct Non-interactive ARgument of Knowledge (zk-SNARK)

To ensure that the data extracted from the TLS session can be validated publicly without revealing sensitive information, our approach will employ zero-knowledge Succinct Non-interactive ARguments of Knowledge (zk-SNARK) (Ben-Sasson, Chiesa, Tromer, and Virza, 2014; Gennaro, Gentry, Parno, and Raykova, 2013; Groth, 2016). For completeness, we will provide an informal definition of zk-SNARK below.

Informally, a zk-SNARK scheme is a cryptographic proof system (Goldwasser, Micali, and Rackoff, 1989) to prove that a public statement, given some private witness, can be decided to be true due to some polynomial-time computable binary relation R. Note that $(\phi, \omega) \in R$ denotes that ϕ is true regarding R given witness ω, and we call ϕ the (public) statement and call ω the (private) witness. Usually, zk-SNARK consists of a triple of algorithms, including Setup, Prover, and Verifier. The Setup algorithm produces a common reference string used throughout the other algorithms. The Prover algorithm, given

the public ϕ and the private ω, produces a proof π, attesting $(\phi, \omega) \in R$. The Verifier, on input the public ϕ and the proof π, returns a bit to reject (0) or accept (1) the proof.

The above zk-SNARK proof system shall satisfy the following security properties:

- *Completeness*: For $(\phi, \omega) \in R$, the prover can always use the witness ω to generate a valid proof regarding the correctness of the statement ϕ, thus passing the verification of the verifier.
- *Zero-Knowledge*: The proof reveals nothing beyond the validity of the statement, i.e., no bit of the private witness ω shall be leaked.
- *Proof-of-Knowledge*: A zk-SNARK satisfies soundness, if the prover cannot efficiently compute a valid proof for $(\phi, \omega) \notin R$. Sometimes, we further need an even stronger soundness property called proof-of-knowledge, that means it is computationally infeasible for any malicious prover to generate a proof without knowing a valid witness ϕ s.t. $(\phi, \omega) \in R$.

In addition to the above security features, zk-SNARK also satisfies:

- *Succinctness*: The proof is short and can be verified quickly (e.g., polylogarithmic in the circuit size of relation R).
- *Non-interactivity*: The proof is generated and published without requiring back-and-forth interaction between the prover and the verifier.

Note that both succinctness and non-interactivity are critical in the blockchain setting (Kosba, Miller, Shi, Wen, and Papamanthou, 2016): succinctness is essential due to the limited on-chain resources available for verifying complex proofs, while non-interactivity is necessary because consensus nodes cannot engage in back-and-forth communication with the prover during the execution of the consensus protocol.

3.5 Two-Party Computation (2PC) with Malicious Security

A two-party computation (2PC) protocol (Beaver, Micali, and Rogaway, 1990; Ben-Or, Goldwasser, and Wigderson, 1988; Goldreich, Micali, and Wigderson, 1987; Yao, 1982) allows two mutually distrustful parties to jointly compute a function over their private inputs while guaranteeing the correctness and privacy of the honest party, even if one of them is corrupted. A 2PC protocol with malicious security (Lindell and Pinkas, 2007; Shelat and Shen, 2011; Wang, Ranellucci, and Katz, 2017) further ensures these guarantees even when the corrupted party may arbitrarily deviate from the prescribed protocol to influence the computation or extract additional information.

However, a fundamental limitation of 2PC protocols lies in their lack of fairness assurance (in the general case) (Cleve, 1986): a malicious participant may obtain the function's output while simultaneously preventing the honest party from learning the result. Our design of a fair TLS data oracle must therefore overcome this inherent limitation of 2PC by leveraging the blockchain and tVDF, thus achieving fairness.

3.6 Abstract of Blockchain

We view the blockchain (which also supports smart contracts) as an ideal bulletin board $\mathcal{F}_{\mathcal{L}}$ (Choudhuri, Green, Jain, Kaptchuk, and Miers, 2017), which also maintains a dictionary $\mathcal{L}$ to track the balance of each user or smart contract. For example, after Alice transfers Ξx to Bob (assuming $\mathcal{L}[\mathsf{Alice}] \geq \Xi x$), then the balance of Alice $\mathcal{L}[\mathsf{Alice}]$ decreases by Ξx and the balance of Bob $\mathcal{L}[\mathsf{Bob}]$ increases by Ξx. Here we let Ξ denote the unit of cryptocurrency on the smart-contract-enabled blockchain.

4 TLS Oracle from "Encrypt-TLS-keys-to-future"

This section presents our protocol for extracting web data from TLS servers in a way that enables the verifier alone to eventually provide the public verifiability of already-transmitted web data. On one side, the design just relies on the non-colluding nature of the prover (i.e., the user client) and the verifier (i.e., the web oracle service provider), instead of assuming a trusted third party (that acts as a designated verifier to honestly forward the extracted TLS data to the public). On the other side, the design also achieves fairness to some extent: once some TLS data was transmitted, a malicious verifier no longer can prevent the prover from selectively disclosing the web data to the public in a verifiable way.

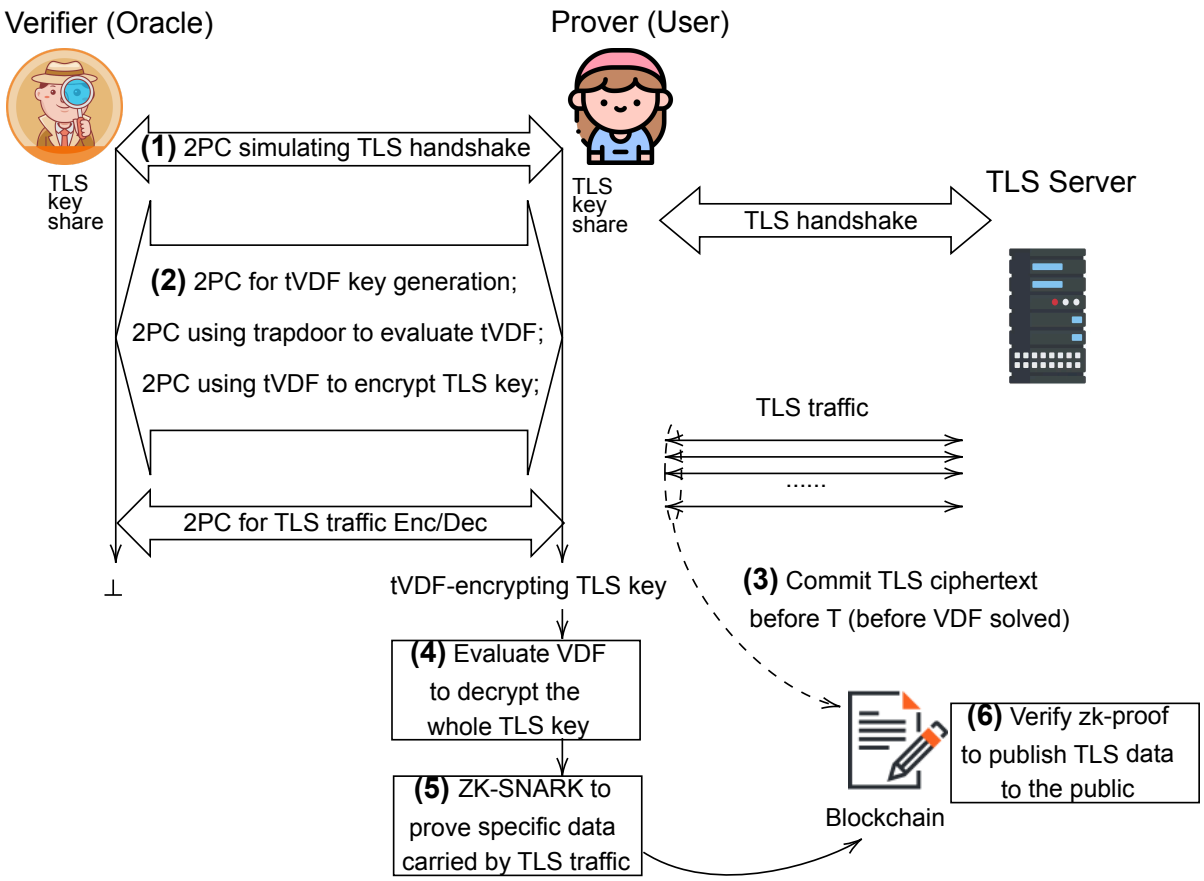


Fig. D1: Execution flow of our TLS oracle protocol with on-chain verifiability and a form of fairness.

4.1 Our TLS Oracle Protocol

The protocol integrates secure two-party computation (2PC) (Wang et al., 2017), a trapdoor verifiable delay function (tVDF) (Wesolowski, 2020), and zero-knowledge proof techniques (Groth, 2016), and is built atop the modern PKI (Cooper et al., 2008; ISO, 2022) and TLS protocols (Dierks and Rescorla, 2008; Rescorla, 2018). The involved parties include: (i) a smart-contract-enabled blockchain; (ii) a website hosting a user's critical data, having a digital certificate and running the standard TLS protocol; (iii) a user as prover and a oracle service provider as verifier, who engage in a collaborative TLS session to emulate a TLS client, thus extracting authentic data from the web server and later collectively disclosing the extracted data on the blockchain.

Protocol Overview. From a very high level, our TLS oracle protocol has an execution flow as Figure D1 illustrates, which proceeds in the following steps:

1. In this phase, the verifier and prover jointly execute a secure two-party computation (2PC) protocol to emulate a TLS client, while the TLS server continues to operate according to the standard TLS specification. At the end of the handshake, neither the prover nor the verifier individually learns the full symmetric TLS session keys. Instead, each party obtains an additive secret share of the session keys, which will later serve as input to subsequent 2PC protocols for AES-GCM encryption and decryption of TLS traffic.
2. Time-lock of TLS key: Next, the verifier and the prover jointly execute another 2PC protocol to perform a sequence of tasks that collectively encrypt the TLS session keys "to the future". These tasks include: (i) the distributed setup of a tVDF, (ii) trapdoor evaluation of the generated tVDF, and (iii) encryption of the TLS session keys using the tVDF output. Since we instantiate the tVDF using primitives over an RSA group, these 2PC computations can be securely realized by adapting existing maliciously secure protocols for secure RSA modulus generation (Burkhardt, Damgård,

Frederiksen, Ghosh, and Orlandi, 2023; Frederiksen, Lindell, Osheter, and Pinkas, 2018; Hazay, Mikkelsen, Rabin, Toft, and Nicolosi, 2019). [92]

3. Commitment of TLS traffic in time: The prover records the transcript of TLS traffic (e.g., AEM-GCM ciphertexts) on the blockchain, before a time that VDF can be feasibly computed without access to the trapdoor.
4. Prover gets TLS key by solving VDF: The prover solves VDF, thus decrypting the TLS session keys.
5. Prover leverages zk-SNARK to prove that specific data is carried in the authenticated TLS ciphertext.
6. The blockchain verifies the zero-knowledge proof, ensuring that specific data is indeed obtained from decryption of the committed TLS ciphertext.

Security Intuition. Informally, the above protocol achieves the following properties (conditioned on that the faithfulness of TLS server and blockchain):

- *Correctness.* If both the verifier and the prover are honest, the authentic data (that is hosted by the TLS server) can be published on the blockchain. This is trivial due to the correctness of all underlying cryptographic schemes.
- *Authenticity against malicious prover.* Despite that the prover could be arbitrarily malicious, the TLS data finally published on the blockchain must be authenticated from the TLS server. This is because the tVDF-encrypting TLS session key is committed to the blockchain (and secretly shared between the verifier and the prover), such that the prover cannot forge authenticated TLS traffic and publish fake TLS traffic on the blockchain before tVDF can be solved.
- *Fairness against malicious verifier.* Since tVDF can finally be solved by the prover, the TLS traffic posted on the blockchain finally can be solely decrypted by the prover (without the involvement of verifier). As such, the prover can always disclose certain TLS traffic (after solving tVDF) if such traffic is indeed transferred from the server to the prover.
- *Private selective disclosure.* The prover can only disclose certain non-sensitive data on the blockchain. This is because (i) the TLS session key is only obtained by the prover after tVDF is solved, and (ii) the zero-knowledge proof from the prover to disclose certain data does not leak additional information.

Efficiency consideration. Efficient instantiation of the above protocol workflow requires addressing several significant efficiency challenges. First, the 2PC protocol for tVDF key generation is highly expensive, as it requires the distributed generation of an RSA modulus (Burkhardt et al., 2023; Frederiksen et al., 2018; Hazay et al., 2019), which is well-known very costly. Moreover, the subsequent 2PC protocol for tVDF's trapdoor evaluation is also expensive.

However, the efficiency of these critical steps can be improved by the following steps: (i) we can enhance tVDF with an extra property of completeness under malicious setup (Chen et al., 2025), such that the prover can always solve tVDF even if the setup of tVDF is done by a malicious party; (ii) then let the verifier to locally run the key generation for tVDF, instead of a 2PC protocol; (iii) finally, we can let the verifier locally run the trapdoor evaluation of tVDF, and that the 2PC protocol for tVDF evaluation can be replaced with a less expensive 2PC protocol for tVDF verification.

4.2 Why not the Alternative Way through Modifying DECO?

One might suggest that our TLS oracle design is not the only possible approach to enabling public verifiability for prior designs such as DECO. In fact, an alternative modification to DECO (Zhang et al., 2020) could enable on-chain verification: instead of committing TLS traffic (Dierks and Rescorla, 2008; Rescorla, 2018) only to the verifier, the prover could commit the ciphertext directly to the blockchain. Then, once the verifier transmits its TLS key share to the prover, the prover can reconstruct the full TLS session keys and prove the correctness of the plaintext corresponding to the committed ciphertext.

However, this approach suffers from a fundamental drawback: it requires the verifier to reliably return and reveal its key share after the TLS communication has completed. This dependency makes the design unsuitable for fair exchange scenarios between fiat and crypto assets, particularly when the verifier is the party transacting fiat currency through the TLS session. If the verifier refuses to reveal

[92] Note that while several optimizations can be applied to improve efficiency of this step, we present a plain version of our oracle protocol for clarity and ease of understanding.

its key share, the prover is unable to complete the proof, even if the TLS event (e.g., payment) did in fact occur. This undermines fairness by enabling strategic aborts, where the verifier can benefit from the transaction while preventing the prover from proving it, thus introducing potential vulnerability in certain applications.

5 Application to Exchange across Web2 and Web3

A TLS oracle protocol with our form of fairness (i.e., the ability to prove certain authenticated data that has already been transmitted, even if the verifier aborts after the transmission) can serve as a critical building block for enabling fair peer-to-peer (P2P) exchange between fiat and cryptocurrency assets, such as between the balance of digital banks and Ethereum tokens (Wood et al., 2014).

Consider a scenario in which a Web3 user, Alice, wishes to exchange her cryptocurrency for fiat currency from a Web2 user, Bob, who completes the fiat payment via a TLS-secured connection to a trusted bank. In traditional protocols, fairness cannot be guaranteed without involving a trusted third party, as Alice could refuse to cooperate in helping Bob prove that he completed the fiat transfer.

In contrast, with our fairness-preserving oracle protocol, Bob is assured that even if Alice aborts the protocol, he can still extract a cryptographic proof of the fiat transaction's completion (e.g., proving that the bank indeed sent a webpage acknowledging a successful transfer). This property eliminates the dependency on Alice's continued cooperation and ensures that the success of the exchange depends solely on the verifiable occurrence of the payment event. Consequently, our protocol establishes a foundation for fair cross-domain asset exchange, bridging Web2 and Web3 ecosystems without reliance on external intermediaries.

Informally, the resulting decentralized P2P exchange protocol attains the following security guarantees:

- *Completeness.* If both Alice and Bob are honest, then Alice receives $\$b$ (i.e., $\Lambda[\mathsf{Alice}]$ increased by $\$b$), and Bob receives Ξa (i.e., $\mathcal{L}[\mathsf{Bob}]$ increased by Ξa).
- *Fairness.* Either (i) Alice receives $\$b$ digital asset and Bob receives Ξa cryptocurrency, or (ii) their cryptocurrency and digital asset balances remain unchanged.
- *Privacy.* Bob's private login credentials (e.g., bank passwords) remain confidential.

For space limit, here we only provide a very brief intuition of our P2P exchange across the Web2 and Web3 domains. Readers interested in further details may refer to our technical paper (Chen et al., 2025).

6 Conclusion and Future Directions

We presented a new TLS oracle protocol that achieves public verifiability and introduces a novel form of fairness, even when one party becomes uncooperative. At its core is the "encrypt-TLS-keys-to-future" paradigm, which allows a prover to reconstruct the TLS session key after a verifiable delay and use zero-knowledge proofs to attest to past TLS communications. Unlike prior approaches such as DECO, our protocol does not rely on a designated verifier remaining online and honest to enable the prover to complete its proof for the extracted TLS data.

To demonstrate the power of our design, we applied it to the problem of cross-domain fair exchange between fiat and crypto currencies. By allowing the fiat holder to (solely) extract the publicly verifiable evidence of a completed bank transaction over TLS, our protocol enables fair peer-to-peer asset exchange without reliance on trusted intermediaries. This construction addresses a long-standing challenge in interoperability between Web2 and Web3 systems.

Our work opens several avenues for future research. First, performance optimizations to the underlying 2PC schemes and other cryptographic protocols could reduce latency and improve scalability, making our paradigm practical for real systems. Second, expanding our model to support more expressive TLS-dependent statements would broaden its applicability beyond currency exchange across domains.

Acknowledgements. Yuan Lu is supported in part by the National Key Research and Development Project of China (2022YFB3102500), the Chinese Academy of Sciences (CAS) Project for Young Scientists in Basic Research (YSBR-035), and the National Natural Science Foundation of China Key Program (No.92270204). Qiang Tang is supported in part by the Australian Research Council (ARC) Discovery Project (DP250101739) and in part by the Research Awards from the Ethereum Foundation, Protocol Laboratories, and the Stellar Development Foundation.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Backman A, Richer J, Sporny M (2024) Http message signatures. Tech. Rep. RFC 9421, Network Working Group

Beaver D, Micali S, Rogaway P (1990) The round complexity of secure protocols. In: Proceedings of the twenty-second annual ACM symposium on Theory of computing, pp 503–513

Ben-Or M, Goldwasser S, Wigderson A (1988) Completeness theorems for non-cryptographic fault-tolerant distributed computation. In: Proceedings of the twentieth annual ACM symposium on Theory of computing, pp 1–10

Ben-Sasson E, Chiesa A, Tromer E, et al (2014) Succinct {Non-Interactive} zero knowledge for a von neumann architecture. In: 23rd USENIX Security Symposium (USENIX Security 14), pp 781–796

Boneh D, Bonneau J, Bünz B, et al (2018) Verifiable delay functions. In: Annual international cryptology conference, Springer, pp 757–788

Burkhardt J, Damgård I, Frederiksen TK, et al (2023) Improved distributed rsa key generation using the miller-rabin test. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, pp 2501–2515

Chen L, Liu Y, Lu Y, et al (2025) Short paper: Decentralized fair exchange for assets across blockchain and the internet. In: 2025 Crypto Valley Conference (CVC), IEEE, pp 102–105

Choudhuri AR, Green M, Jain A, et al (2017) Fairness in an unfair world: Fair multiparty computation from public bulletin boards. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp 719–728

Cleve R (1986) Limits on the security of coin flips when half the processors are faulty. In: Proceedings of the eighteenth annual ACM symposium on Theory of computing, pp 364–369

Cooper D, Santesson S, Farrell S, et al (2008) Internet x. 509 public key infrastructure certificate and certificate revocation list (crl) profile. Tech. rep.

Dierks T, Rescorla E (2008) The transport layer security (tls) protocol version 1.2. Tech. Rep. RFC 5246, Network Working Group

Dworkin M (2007) Recommendation for block cipher modes of operation: Galois/Counter Mode (GCM) and GMAC. Tech. Rep. NIST Special Publication 800-38D, National Institute of Standards and Technology (NIST)

Dworkin MJ, Barker E, Nechvatal JR, et al (2001) Advanced encryption standard

Frederiksen TK, Lindell Y, Osheter V, et al (2018) Fast distributed rsa key generation for semi-honest and malicious adversaries. In: Annual International Cryptology Conference, Springer, pp 331–361

Gennaro R, Gentry C, Parno B, et al (2013) Quadratic span programs and succinct nizks without pcps. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, Springer, pp 626–645

Goldreich O, Micali S, Wigderson A (1987) How to play any mental game. In: Proceedings of the nineteenth annual ACM symposium on Theory of computing, pp 218–229

Goldwasser S, Micali S, Rackoff C (1989) The knowledge complexity of interactive proof systems. SIAM Journal on Computing 18(1):186–208

Groth J (2016) On the size of pairing-based non-interactive arguments. In: Annual international conference on the theory and applications of cryptographic techniques, Springer, pp 305–326

Hazay C, Mikkelsen GL, Rabin T, et al (2019) Efficient rsa key generation and threshold paillier in the two-party setting. Journal of Cryptology 32(2):265–323

ISO (2022) Information technology — public key infrastructure — practices and policy framework. Tech. Rep. ISO/IEC 27099:2022

Kosba A, Miller A, Shi E, et al (2016) Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. In: 2016 IEEE symposium on security and privacy (SP), IEEE, pp 839–858

Krawczyk H, Bellare M, Canetti R (1997) Hmac: Keyed-hashing for message authentication. Tech. Rep. RFC 2104, Network Working Group

Labs P (2017) Filecoin: A decentralized storage network. Tech. rep., URL https://filecoin.io/filecoin.pdf, whitepaper

Lauinger J, Ernstberger J, Finkenzeller A, et al (2025) Janus: Fast privacy-preserving data provenance for tls. Proceedings on Privacy Enhancing Technologies

Lindell Y, Pinkas B (2007) An efficient protocol for secure two-party computation in the presence of malicious adversaries. In: Annual international conference on the theory and applications of cryptographic techniques, Springer, pp 52–78

McGrew D (2008) An interface and algorithms for authenticated encryption. Tech. Rep. RFC 5116, Network Working Group

Nakamoto S (2008) Bitcoin: A peer-to-peer electronic cash system. Available at SSRN 3440802

Nir Y, Josefsson S, Pégourié-Gonnard M (2018) Elliptic curve cryptography (ecc) cipher suites for transport layer security (tls) versions 1.2 and earlier. Tech. rep.

Pietrzak K (2019) Simple verifiable delay functions. In: 10th innovations in theoretical computer science conference (itcs 2019), Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pp 60–1

Qualys (????) SSL Pulse. https://www.ssllabs.com/ssl-pulse/, accessed: 2025-2-28

Rescorla E (2000) Http over tls. Tech. Rep. RFC 2818, Network Working Group

Rescorla E (2018) The transport layer security (tls) protocol version 1.3. Tech. Rep. RFC 8446, Internet Engineering Task Force (IETF)

Rogaway P (2002) Authenticated-encryption with associated-data. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp 98–107

Salowey J, Choudhury A, McGrew D (2008) Aes galois counter mode (gcm) cipher suites for tls. Tech. Rep. RFC 5288, Network Working Group

Shelat A, Shen CH (2011) Two-output secure computation with malicious adversaries. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, Springer, pp 386–405

Wang X, Ranellucci S, Katz J (2017) Authenticated garbling and efficient maliciously secure two-party computation. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp 21–37

Wesolowski B (2020) Efficient verifiable delay functions. Journal of Cryptology 33(4):2113–2147

Wood G, et al (2014) Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper 151(2014):1–32

Xie X, Yang K, Wang X, et al (2024) Lightweight authentication of web data via garble-then-prove. In: Proceedings of the 33rd USENIX Conference on Security Symposium, pp 1957–1974

Yao AC (1982) Protocols for secure computations. In: 23rd annual symposium on foundations of computer science (sfcs 1982), IEEE, pp 160–164

Zhang F, Cecchetti E, Croman K, et al (2016) Town crier: An authenticated data feed for smart contracts. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 270–282

Zhang F, Maram D, Malvai H, et al (2020) Deco: Liberating web data using decentralized oracles for tls. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp 1919–1938

Rebuilding Social Cohesion: AI & The Faces of Trust

Thomas Souverain[1]

[1] Department of Philosophy & AI Ethics, French Alternative Energies and Atomic Energy Commission (CEA), Paris-Saclay, France.

Corresponding author. E-mail: thomas.souverain@cea.fr

Abstract

This paper is dedicated to introduce the fourth thematic of the *Paris Journal on AI & Digital Ethics*: Rebuilding social cohesion in AI-powered societies. We situate the full papers of the track in a broader context: what do they tell us about *trust* in the era of AI? The Introduction 1 opens with a paradox concerning trust, between the presence of the word and the absence of the plain phenomenon. If trust is called upon, it is that the notion behind needs to be restored. We then distinguish between two types of trust, depending on the considered *scale*. At the political or social level, AI tools play a classical dual role to promote participatory governance or increase the opacity of decisions. At the individual level, AI bots and recommendations can be treated as trusted third parties, when effectively interacting with people. As applying interpersonal trusting effects to AI models is not obvious, we discuss briefly in Section 2 how moral, affective and cognitive mechanisms might be mirrored from human to machine trustees. Through our lens on trust, we finally propose a view on the papers of the track. From philosophical theories to empirical studies, from social to cognitive frameworks, we argue in Section 3 that they complete each other to picture the new landscape of social trust mediated by AI, and individual trust in AI devices.

Keywords

Trust, AI, Philosophy, Affective, Moral.

1 Introduction: Trust in AI, as good as gold

Trust could be the sly genie of a tale, or the subtle answer of a riddle. Once we begin to invoke it, it is already lacking ; it just disappeared.

Noteworthy is trust at the forefront of mediatic discussions on production and governance of Artificial Intelligence (AI). For instance, the recent European legislation on AI systems set as its purpose 'to promote the uptake of human-centric and trustworthy AI', in order 'to improve the functioning of the internal market' [93].

In fact, the new wave of connectionnist AI models, which can learn patterns from data without being fully dependent on human instructions to do so, earned the nickname of 'black boxes' among AI researchers (Gunning and Aha, 2019). The opacity of such predictive models, which increased in precision in the 2010s from borrowers' scoring to text generation, naturally fed a traditional critique against the use of technology as a leverage for power and surveilance (Foucault, 2004). The mediation introduced by technological devices digs a gap between designers and lay users. This feeds a natural suspicion on the decisions behind their settings. As has been criticized on the government of experts in indirect democracy (Ruphy, 2019), the design and decision procedures would happen too far from lay citizens and would be too complex for them. Hence, the predictions or generations of AI would be easily manipulable and oriented towards the preferences of their benefactors (O'Neil, 2017).

In this context, we can understand the frequent call for trust as the exact symmetric for the distrust in ruling elites felt by citizens. There is an urgent need for the genie; yet, the lamp has been swamped by sand dunes of decisions, distant from the subjects, translated in the Elamite letterings of computer science. We hypothesize that governance, communication, or regulation initiatives try to plug the gaps of distrust in ruling bodies, with afterwards initiatives on AI devices.

We wish then to complete the charade and compare trust, *when it is present*, with money. So long as its power remains valid, it is ordinarily used for trades, such that noone even questions the value behind gold or banknote standards (Simmel, Frisby, Bottomore, and Lemert, 2011). Samewise, in technical designs, trust is demonstrated as effective when the banknotes of AI bots and predictions are used without wariness, on an everyday basis. In our view, the risk-based approach of the European regulation precisely wants to tackle this problem. The central role of standards to presume conformity of AI providers and deployers [94], points towards a direction where AI devices would be used as other technical tools regardless of their complexity, such as computers and elevators.

2 Multiple dimensions of trust with AI : scales and acceptance mechanisms

In that regard, AI models used to recommend, to filter or to synthesize contents reconduct a classical problem of *dual use* in philosophy of techniques (Forge, 2010). Recommender systems can present information strengthening the beliefs of their users, or broadening their view ; Large Language Models can mount or pacify debates, blurring the lines of truth (Spitale, Biller-Andorno, and Germani, 2023).

Nonetheless, AI is not only impacting trust as a supplementary tool mediating deciders and citizens ; AI is also an object of trust. In the following Section, we clarify the different extents to which current AI techniques affect trust. Whereas the social trust, as *cohesion* between humans, partly depends on the moral use of AI mediations, we further investigate how philosophical accounts on inter-human trust might be extended to human interactions with AI operators. It might indeed seem puzzling for institutions to target a precise goal, in order to restore a favorable climate to exchanges. Trust in the era of AI opens multiple directions and especially, two types of questions.

- **Scales of trust, from social to interpersonal connection**

  The exact scale of the trust to restore is hard to isolate, as in the way the problem emerges, designers and design are entangled. To build trust in the era of modern AI, shall we focus on trust in humans designing AI, and more generally the rationality behind rules and institutions? Alternatively, do we need to increase trust in particular AI systems deployed in 'the internal market' in terms of the European Commission (UE, 2024), with whom people interact? These are two objects which evolve pairwise, yet, deserve separate study in the way they engender trust.

[93] This objective is invoked at the opening of the European legal text. Cf. (UE, 2024), preamble, recital 1.
[94] Cf. (UE, 2024) standardization request to national AI industrial actors, Article 40.

- **Dimensions of human-AI trust**

  What is exactly (dis)trust, that it can be felt? Since the 1980s, philosophical discussions have emerged on the nature of trust, mostly through the lens of human interpersonal trust. The majority supports trust to be more than strict "reliance" (Hawley, 2014), reliance amounting to the acceptance of the trustor to depend on an external agent in a context of uncertainty.

Yet, the 'je-ne-sais-quoi' added to reliance does not lead to consensus. Three major directions to characterize trust are identified by Simpson (2012): conative, affective, and cognitive. Extending these three interpersonal accounts of trust to human trust in AI systems is not obvious:

2.1 Moral trust: feeling gratitude or betrayal from AI

A conative or moral account of trust inherits from the Kantian distinction between a person and a thing [95]. I will consider a person as a moral subject, which is not reducible to an object or a mean to achieve my own goals. This confers rights to this entity, being an end in oneself, but also specific duties. Especially, I might have moral expectations regarding this person, whose reasonable nature intimates she won't break promises. In particular, the 'participant stance' of Holton (1994) illustrates how moral agents are implicitly tied, from the point of view of the trustor:

"When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld. [...]" [96]

Of interest for us is the mention of "someone". If I expect a table to hold a Ming vase while I go out to pick some lilies, and the vase breaks because of a window suddenly opening, I won't "feel betrayal" in Holton's view. At most will I be angry against my rational anticipation which did not take into account the window. Let us replace the table by Mr Dupond, whose mission would be to hold the piece of porcelain while I walk to the other room. If the vase falls, my angriness against Mr Dupond will become a moral feeling. I will feel "betrayed" because I had a moral idea of his trustworthiness, which broke up with the rare vessel.

2.2 Affective trust: dependencies to AI companions

Transferring this view on interpersonal trust to human-AI trust raises problems. From an ontological point of view, a statistical model delivering scores of creditworthiness, or assembling coherent sentences, is not ensured to be a moral subject. However, regarding trust as an *effect*, the question of what *is* an AI system might be put aside, when considering what *happens* when confronted to AI operations. In that direction, empirical studies of AI deployed in medical, recruiting or scientific research contexts were associated with implicit moral considerations of their users (Ferrario and Loi, 2022; Sivaraman, Bukowski, Levin, Kahn, and Perer, 2023; Souverain, 2025). While the extreme reaction of betrayal was hard to locate, negative considerations resembling more to angriness in cases of bugs, gratitude emerged to qualify smart and time-saving recommendations of the "little robots" (Souverain, 2025).

This leads us to another trait of trust analysed by philosophers, which involves affective dependency. This point joins psychological and sociological studies, who show that AI devices used on a daily basis upreared some attachment (Velkovska and Relieu, 2020). This attachment was shown for lay users, but also for domain experts. Despite they rationally know AI systems to be their work devices, they can express affection for the system, employing diminutives and revealing emotions in front of their outputs (Sivaraman et al., 2023). Though Baier (1986) illustrates affective dependency with the position of a baby regarding one's mother, we do not see such extreme vulnerability as necessary. The exposition to the reaction of another person, considered with feelings, seems us more realistic in everyday occurrences of 'trust', hence, sufficient to analyse the affective component of trust in humans and in machines.

2.3 Cognitive trust: confident rational anticipations of AI performances

The last type of definition of trust as a supplement to mere reliance is "cognitive" for Simpson (2012). O'neill (2002) illustrates this with a patient which has been going to see the same doctor for years, even if she knows that the doctor feels irritation about her. In that situation, she still accepts dependency on the treatments prescribed by the medicine, out of any emotional link. Why? This is because she rationally expects that the doctor will do his job.

[95] Cf. Kant (2020), chapter 2, pp. 28-29.
[96] Cf. Holton (1994), p. 67.

An approach such as Hardin (2002) on "encapsulated interests," drawing on economic models of maximizing rationality, makes explicit the rationale behind cognitive trust. It is most plausible that the patient believes the surgeon is motivated to provide the best possible care—whether for the sake of her reputation, her remuneration, or to avoid a lawsuit. Trust, in this sense, would be expressed as the equivalence between "I believe you are trustworthy" and "I trust you" ((Hardin, 2002), p. 10).

As for the affective account, we do not see the estimation of the trustworthiness of an agent as a necessity for cognitive trust. What Simpson (2012) calls "predictive trust" shares the same projection,that is, the rational anticipation of an event. To put it in computational terms, if my rational expectations that an AI chatbot will indicate truth on some political or historical fact (which can be statistically founded) exceed a confidence threshold, my mental probabilities on the reliability of this AI justify the presence of cognitive trust. This mental model is indeed added to the reliance or acceptance of the models' predictions, or founds it.

3 Track on the dual role of AI facing trust and social cohesion

AI systems can affect the two types of trust described above, political and interpersonal. The following track of the *Paris Journal on AI & Digital Ethics* provides original analysis of the role played by AI to move the boundaries of trust. To picture the various colors of this remodeled trust, we gather contributions from philosophy, law, social sciences, linguistics, and human-computer interactions.

On a political side, AI adds a further layer of technological mediation between citizens and their representatives, or between consumers and tech leaders. Through the juridical lens of prompts' copyrights, the first paper (Simas C. de Oliveira, 2025) shows how instant it is to recognize the original works of humans with generative AI. Institutions face a challenge of crediting innovative and artistic works as trustworthy, measuring the extent to which generative AI contributed. AI new technological mediation might also dig the gap felt by people which stand far away from technologies and decision processes. To identify and address this issue, Bougnères, Besançon, and Basol (2025) provide us with a unique experiment among 630 disadvantaged young adults, linking emotional perceptions of trust in institutions with the ability to distinguish opinions and facts.

Digging into political life more generally, the philosophical paper of Tallent (2025) suggests that AI can serve as a tool or a substitute for democracy. At this scale, trust would amount to social cohesion between citizens. AI recommendations could foster it, e.g. giving equitable places to diverse opinions and pacifying debates. The main challenge would be, instead of overly smoothing contradictory points of view (as the author pinpoints with the Habermas machine of Google researchers), to set governance with online communications and recommendation algorithms to go along with in-depth discussions (Tessler, Bakker, Jarrett, Sheahan, Chadwick, Koster, Evans, Campbell-Gillingham, Collins, Parkes et al., 2024). Bridges between political cohesion and trust of individuals in AI devices might be built, when considering the rationality behind expectations of AI outputs (see Section 2.3). In her talk given during the Conference, Prunkl (2025) drew a philosophical parallel between the ruled process of algorithms and the modern process of bureaucratisation, both enabling trust in the supposed neutrality of decisions by the absence of arbitrary interventions.

Yet, humans do not only praise and blame their pairs from a rational perspective. They are not only handling with AI outputs with analytical and expectative judgements. At the same time, they are connecting with AI systems through moral and emotional stances, which must also be taken into account when studying trust in AI.

The two last papers supply the track in that direction. They introduce in-depth analysis of AI systems, regarding how AI devices entwine rationality and emotional attachments of humans. Markelius, Chaves, Spencer, and Kuiper (2025) warn for the risk of irrational attachments to "Companion AI", as chatbots whose goal is to foster social and emotional engagement of humans. Through flattering mechanisms or corporate settings to maximize involvement, generative AI embedded into the platforms promote the affective component of trust over rational or moral contracts.

At the opposite side of the spectrum, Upton Patton, Kleiner, Miller, Deas, Grieser, Shepard, Vente, and McKeown (2025) explore the ways AI recommendations might be a tool to enhance social connection, fostering constructive and positive interactions on social media. They launch an experimental platform where New York students from minorities shared feelings and thoughts, and showed that AI recommenders had the power to amplify joy and foster empathic interactions online.

4 Conclusion

All together, the rich interdisciplinary studies gathered here highlight that the boundaries of trust are changing due to the quick spread of AI technologies.

Everyday uses of generative AI and recommender systems mediate our social interactions, with the risk of emptying the notions of political debate and creativity. At the same time, they present an opportunity to change the dividing-lines of representation in political life, and to deeply re-interrogate the human specificities of creation and trust.

We can not, however, load the shoulders of states with the whole burden. Designing companies, especially the most influent ones, are responsible for setting objectives which reduce the *affective* dependency triggered by AI operators. Dealing with AI words, images, recommendations and predictions, *rational expectations* of AI outputs should be the main basis to establish trustworthiness, and consequently trust.

Our synthesis of these papers calls for a pairwise work of democratic institutions and business objectives, serving two aspirations : to target a proper trust in AI systems, anchored in rational assessment ; to orientate the use of AI towards public purposes in long-term, fair, and participatory ways.

Acknowledgments

This article delves into the notion of trust, connecting within each other the full papers of abstracts presented on 2025, June 17th, at the Paris Conference on AI & Digital Ethics, hosted by Sorbonne University. We thank Dr. Hubert Etienne for having put forward this initiative four years ago, opening an unshallow interdisciplinary dialogue on the digitalization of our world.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Baier A (1986) Trust and antitrust. ethics 96(2):231–260

Bougnères A, Besançon H, Basol M (2025) Critical thinking and media literacy in youth from underserved contexts: Evidence from a field intervention. The Paris Journal on AI & Digital Ethics

Ferrario A, Loi M (2022) How explainability contributes to trust in ai. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 1457–1466

Forge J (2010) A note on the definition of "dual use". Science and Engineering Ethics 16(1):111–118

Foucault M (2004) Naissance de la biopolitique: cours au Collège de France (1978-1979)

Gunning D, Aha D (2019) Darpa's explainable artificial intelligence (xai) program. AI magazine 40(2):44–58

Hardin R (2002) Trust and trustworthiness. Russell Sage Foundation

Hawley K (2014) Trust, distrust and commitment. Noûs 48(1):1–20

Holton R (1994) Deciding to trust, coming to believe. Australasian journal of philosophy 72(1):63–76

Kant I (2020) Groundwork of the metaphysic of morals. In: Immanuel Kant. Routledge, p 17–98, URL https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf

Markelius A, Chaves P, Spencer SW, et al (2025) The trust paradox: How companion ai is rewiring human connection and social cohesion. The Paris Journal on AI & Digital Ethics

Simas C. de Oliveira J (2025) Who owns the words? copyright, ai prompts, and the governance of digital authorship. The Paris Journal on AI & Digital Ethics

O'Neil C (2017) Weapons of math destruction: How big data increases inequality and threatens democracy. Crown

O'neill O (2002) Autonomy and trust in bioethics. Cambridge University Press

Prunkl C (2025) Ai as bureaucratic decision-making. The Paris Journal on AI & Digital Ethics

Ruphy S (2019) Public participation in the setting of research and innovation agenda: Virtues and challenges from a philosophical perspective. In: Innovation Beyond Technology: Science for Society and Interdisciplinary Approaches. Springer, p 243–263

Simmel G, Frisby D, Bottomore T, et al (2011) The philosophy of money. Routledge

Simpson TW (2012) What is trust? Pacific Philosophical Quarterly 93(4):550–569

Sivaraman V, Bukowski LA, Levin J, et al (2023) Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp 1–18

Souverain T (2025) Ai to renew public employment services? explanation and trust of domain experts. AI and Ethics 5(1):55–70

Spitale G, Biller-Andorno N, Germani F (2023) Ai model gpt-3 (dis) informs us better than humans. Science Advances 9(26):eadh1850

Tallent A (2025) Ai, digital democracy, and the crisis of social cohesion: Rebuilding political trust. The Paris Journal on AI & Digital Ethics

Tessler MH, Bakker MA, Jarrett D, et al (2024) Ai can help humans find common ground in democratic deliberation. Science 386(6719):eadq2852

UE (2024) Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Upton Patton D, Kleiner S, Miller S, et al (2025) Ai with heart: A joy-centered framework for ethical ai development implementation. The Paris Journal on AI & Digital Ethics

Velkovska J, Relieu M (2020) Pourquoi ethnographier les interactions avec les agents conversationnels? Réseaux (2):9–20

Who Owns the Words? Copyright, AI Prompts, and the Governance of Digital Authorship

Jaqueline Simas C. de Oliveira[1]

[1] Law Graduate Program (PPGD), Fundação Getulio Vargas Rio (FGV-Rio), Praia de Botafogo, 190, Botafogo, Rio de Janeiro, RJ, Brazil.

Corresponding author. E-mail: oliveira.jaqueline@fgv.edu.br

Abstract

As generative artificial intelligence becomes increasingly embedded in creative and professional workflows, legal frameworks face new challenges in defining authorship and protecting human contributions. While much of the legal debate has focused on the outputs of GenAI systems, this paper turns attention to the prompt — textual input crafted by users to direct AI models. Some prompts require substantial creativity, effort, and stylistic intention, raising the question of whether they qualify as copyrightable works.

Focusing on the Brazilian Copyright Act, which requires originality for protection, this paper argues that certain prompts may satisfy this criterion, particularly those that reflect narrative vision, artistic direction, or aesthetic expression. Through a normative legal analysis, complemented by comparative references and illustrative examples, the paper outlines how current Brazilian law could accommodate the protection of creative prompts without legislative reform. It also explores the boundaries of this protection, drawing attention to the functional limits imposed by doctrines such as the merger doctrine.

Beyond doctrinal considerations, the paper situates the issue within broader concerns about governance and digital trust. Recognizing the legal value of prompts may enhance user agency, clarify ownership in AI-mediated processes, and foster more inclusive and transparent ecosystems.

Keywords

Copyright Law, Prompts, Brazilian Law, Generative AI.

1 Introduction

Regulatory debates involving new technologies and the agents associated with them have intensified in recent decades, especially from milestones that have had a horizontal impact on society, such as the popularization of the internet in the early 2000s and, more recently, the spread of the use of generative artificial intelligence (AI), which, until recently, seemed a distant and futuristic reality.

The law continually seeks answers to new questions that arise from changing contexts. These are issues that challenge almost all legal fields: civil liability, data privacy, intellectual property, taxation, and criminal law are just a few examples. The legal field is gradually shifting from treating "digital law" as a distinct and isolated area of study toward integrating digitalization into traditional domains, prompting a reinterpretation of established legal frameworks.

The use of generative AI became popular during the pandemic with ChatGPT, OpenAI's AI model designed to process and generate text in a coherent and contextual way, and has since diversified enormously. In a short space of time, the possible applications of generative AI have become countless. Not only has ChatGPT been improved, but it has gained direct competitors (such as Google's Gemini and Meta's Meta AI) and paved the way for the development of millions of other applications (whether based on OpenAI's language models or new models created). These new applications include, for example, tools that create advertising texts (such as Jasper), new music (such as Amper Music), or assist with legal tasks (such as Harvey).

The results (or "*outputs*") of a generative AI tool vary considerably according to the *"prompts"* of its users. In this paper, *prompts* are the instructions with textual content used by the user of generative AI with the objective that generative AI produces new content, whether textual, audiovisual, musical, etc. While every *prompt* is, in the context of AI, an input, not every input is a *prompt*. Inputs are all the elements that power generative AI so that it can process and generate new results. Thus, they include, in addition to *prompts*, the database used by developers and programmers for training the AI model, for example.

The impact of *prompts* on the results generated by generative AIs is indisputable. The *prompt* can dictate, among many other variables, the extent of the content to be generated, the tone to be employed, and the emphasis to be given. Thus, different prompts will cause generative AI to return different results, even when based on a similar or even the same idea.

Prompts vary considerably in complexity. Some are simple directives, such as "write a poem about the ocean," while others resemble detailed scripts, stylistic blueprints, or artistic visions crafted to elicit precise effects. Despite the fact that these prompts often involve substantial creative input from users, their legal status remains ambiguous.

Much has been discussed about the challenges of legal protection of generative AI results (Schirru, 2020; Rocha de Souza, Schirru, Lana, and Ramos, 2024)[97] This debate is, without a doubt, extremely relevant. In Brazil – the primary focus of this paper -, the legislation in force is expressed in not allowing copyright protection of results generated by AI, [98] a fact that denotes greater urgency in the dialogue on the need for regulatory updates in copyright rules.

Nevertheless, the expressive and creative aspects inherent in complex prompts - such as those found in visual arts, music, or storytelling - highlight the necessity for the debate to begin with, or at least encompass, an analysis of the legal protection afforded to prompts and the potential implications that such protection, or its absence, may entail. After all, the results generated by generative AIs are only possible from the *prompts*, although other elements are also crucial (*e.g.,* the variety of the database used to train the AI, the use of algorithms that guide the analysis and decision-making of the AI, etc.). In addition, any understanding of the legal protection of *prompts* may have direct consequences on the analysis of the protection of the outputs.

Although the specific theme of copyright protection of *prompts* has already been a more in-depth object of foreign academic literature (Lee, 2023; Mazzi, 2024), intellectual production from the perspective of Brazilian legislation is still scarce. [99]Even if this were not the case, the fact is that the current

[97]From the perspective of foreign jurisprudence, see, for example, Thaler v. Perlmutter, No. 22-cv-1564, ECF No. 24 (D.D.C. Aug. 18, 2023), US; Decision of the US Copyright Office in the case "Théatre D'opéra Spatial (SR # 1-11743923581)" and Hamburg Regional Court, Germany [2024]: Robert Kneschke v. LAION e.V., Case No. 310 O 227/23, Germany. From a legislative perspective, see, for example, the EU IA Act, already approved in the European Union, and Bill No. 2,338,2023 under discussion in the Brazilian Congress.

[98]Law No. 9,610/98 (or Brazilian Copyright Act) is expressed in the sense that "author" is the natural person who created the authorial work (art. 11).

[99]There is already relevant academic production in Brazil on the protection of content generated by generative AI. However, specific analyses on prompt protection seem to be limited to opinion articles, such as, for example, STIGERT, K. Copyright

moment requires academic production to foster issues and contribute to the theme and regulatory challenges imposed by new technologies in traditional fields of law, as in the case of generative AI with intellectual property, the area of protection and study of copyright. This need is even more pressing in the Brazilian scenario, given the mismatch between the date of entry into force of Law No. 9,610 of 1998 ("Brazilian Copyright Act"), and the beginning of discussions involving copyright protection of elements related to generative AI, such as prompts and their results.

Given the above scenario, the focus of this paper is to analyze whether the *prompts* can be considered copyrighted works protected by the Brazilian Copyright Act in Brazil (1) and, if so, what the extent and limitations of this protection (2). A secondary goal is to outline the potential legal and social effects of acknowledging or not acknowledging prompt authorship, particularly with respect to regulation, user rights, trust, and digital authorship governance.

With these objectives in mind, **Section 1** of this paper proposes to review the originality criterion of the works, which is established by the current legislation as one of the requirements for a given work to be subject to copyright protection[100] and to examine the intersection between these requirements and the *prompts*, to ascertain the appropriateness of copyright protection, emphasizing the extent and limitations of protection.

Section 2 of this document examines the legal and social implications associated with granting copyright protection to prompts, including examples such as disputes over prompt ownership, implications for user rights and potential impacts on the governance of digital authorship. Rather than providing an exhaustive analysis, Section 2 aims to highlight selected consequences arising from the findings in Section 1, to foster further research and academic debate on the regulation of generative AI.

Finally, **Section 3** summarizes the main conclusions of this study, the main one being the idea that prompts are subject to copyright protection under current legislation, but that this protection is quite limited and depends mainly on a case-by-case analysis. In this sense, it is concluded that certain prompts are mere instructions or questions that cannot be considered "creations of the spirit" subject to protection by the Brazilian Copyright Act. In these cases, by analogy, the prompts would be equated to the guidelines for the creation of an authorial work given by an employer to its employee or by a contractor to an artist of a commissioned work. Incidentally, it is concluded that the recognition of the copyright protection of *prompts* will generate numerous legal and social consequences that will require great reflection from academia and regulators.

This paper adopts a dual-method approach. First, it conducts a doctrinal legal analysis of the Brazilian Copyright Act, examining statutory provisions, case law, and doctrinal commentary. Second, it incorporates an interdisciplinary perspective, drawing on literature from digital governance, human–AI interaction, and comparative law. Sources were selected based on relevance, recency, and their contribution to the debate on digital authorship and AI. Comparative cases were chosen for their influence on international copyright discourse and their potential analogies to the Brazilian context.

Although this paper focuses on Brazilian law, it refers to U.S. cases and policy documents as comparative heuristics for a single reason: both systems require "originality", yet they operationalize it differently. As it will be exposed, the U.S. doctrine articulates a low but explicit threshold and, in recent AI discussions, often stresses the unpredictability of outputs to deny protection. Brazil, anchored in the civil-law *droit d'auteur* tradition, [101] centers analysis on the author's personal creative expression (*"criação do espírito*) and moral rights. Citing U.S. materials thus serve to illuminate, not to transpose, and the analogies are drawn only insofar as they clarify how Brazilian law can assess the originality of prompts.

While this paper focuses on the Brazilian context, the issues addressed are increasingly relevant in international debates, particularly in jurisdictions influenced by the *droit d'auteur* tradition.

---

of *Artificial Intelligence Prompts* - An analysis of the "Théâtre d'Opéra Spatial" case. Jusbrasil, 2023. Available at this URL. Accessed on: Jan. 6. 2025

[100] This is not the only criteria for copyright protection provided for by the LDA. Other criteria include, for example, the fixation of the work and the creation of the work by an **individual** within 70 years from his death. However, for the purposes of the discussion involving *prompts*, the requirement of originality is the most relevant to be considered, given its great subjectivity.

[101] While the U.S legal system is based on the copyright tradition, Brazil belongs to the civil-law droit d'auteur tradition, which places strong emphasis on the author's personal creative expression and moral rights (e.g. attribution and integrity). While originality exists in both copyright and droit d'auteur traditions, *droit d'auteur* frames it as a manifestation f the author's personality-ofter captured in Brazil by the notion of *"criações de espírito"*. This baseline shapes the analysis here and clarifies why a focus on the input (the prompt as an expressive textual work) is doctrinally coherent in Brazil, even when U.S. practice tends to focus on output unpredictability

2 The Brazilian Copyright Act and its applicability to prompts

The analysis of the protection of *prompts* by copyright in Brazil necessarily involves the review of the criterion of originality, a requirement for legal recognition. Although other requirements are also imposed by Brazilian legislation, the requirement of originality, in addition to being seminal for the recognition of the work, is the most subjective among them. Thus, the assessment of whether *or not a prompt* can be considered original is the gateway to any discussion involving the legal protection of *prompts* by the Brazilian Copyright Act. The fulfillment of other requirements (such as the fixation on tangible or intangible support or the time limit of protection) is easier to verify. This Section 1, therefore, will make a general review of the criterion of originality in the field of copyright (2.1.) and, subsequently, examine its application to prompts, also highlighting the limits of possible protection (2.3.).

2.1 Revision of the originality criterion established by the Brazilian Copyright Act

Copyright protection in Brazil, in addition to being recognized by the Federal Constitution[102], is mainly regulated by the Brazilian Copyright Act. Computer programs are also considered works subject to protection by copyright legislation (Section 7, XII of the Brazilian Copyright Act), but, given their particularities, their protection is complemented by Law No. 9,609/98 (or Brazilian Software Act). Both laws resulted from the adaptation of the national intellectual property system to the Agreement on Trade-Related Aspects of Intellectual Property Rights ("TRIPS"), to which Brazil is a signatory. As defined in the Introduction, this paper will focus solely on the Brazilian Copyright Act, as the originality criteria applicable to human-authored textual expressions - such as prompts - is defined by it.

According to the Brazilian Copyright Act, *"[t]he creations of the spirit, expressed by any means or fixed in any support, tangible or intangible, known or invented in the future, are protected intellectual works*" (Section 7), which must have been created by a natural person, the author (Section 11). Although the requirement of originality is not expressly defined by the Brazilian Copyright Act (and it is not expressly defined by the Berne Convention, an international treaty of 1886 on copyright, to which Brazil has been a signatory since 1922 and which was partially received by TRIPS), based on the definition of its Section 7 and the entire history of the evolution of copyright, it has been agreed upon, in caselaw (Tribunal de Justiça de São Paulo, 2014; Tribunal de Justiça do Rio Grande do Sul, 2017) and doctrine (Bittar, 1992; Paranaguá and Branco, 2009) that the copyrighted work subject to protection is the one endowed with originality that has been fixed (i.e., materialized)[103] in tangible or intangible support.[104]

The criterion of originality imposes a great challenge of analysis, because of its subjectivity (Barbosa, 2013). The Brazilian Copyright Act itself, in proposing its scope of protection mentioned above, is excessively comprehensive and vague. The analysis of originality will result from a value judgment on whether or not a certain element is a "creation of the spirit". Thus, originality requires that the work has differentiating elements added by the author (i.e., his creations of spirit) in relation to everything that already exists, including other pre-existing works (Paranaguá and Branco, 2009; U.S. Supreme Court, 1991). To be considered original, the work does not necessarily have to be entirely new. It may even have elements of previous authorial works, as long as the modifications made are endowed with originality, that is, they result in a new expression. These situations qualify as derivative works under Section 5, VIII, g of the Brazilian Copyright Act. [105]

The Brazilian Superior Court of Justice ("STJ"), in a case that dealt with copyright on musical works, summarized the condition for compliance with the requirement of originality as follows: "(...) *For the work to enjoy protection, it is required that it be distinct from the pre-existing ones, and not be limited to a mere copy or appropriation of elements found in other works*" (Superior Tribunal de Justiça, 2020).

As the criterion of originality for copyright protection is shared by several jurisdictions around the world, it is also worth highlighting the paradigmatic *Feist case* judged by the Supreme Court of the United States, which established that *"a) Article I, § 8, cl. 8, of the Constitution mandates originality*

[102] Section 5, XXVII and XXVIII.

[103] Some authors, such as Barbosa (2013), differentiate fixation from materialization, in the sense that materialization would be prior to fixation and independent of the support (for example, music would be materialized when it was sung, regardless of whether it was fixed by means of a recording).

[104] These are not the only criteria for copyright protection provided for by the Brazilian Copyright Act. Other criteria include, for example, the creation of the work by an individual and that it has been created within 70 years from the date of its author. However, for the purposes of the discussion involving *prompts*, these are the requirements that represent greater complexities to consider.

[105] The creation of derivative works depends on the authorization of the author of the original work.

as a prerequisite for copyright protection. The constitutional requirement necessitates independent creation plus a modicum of creativity.," concluding that "[o]*riginality requires only that the author make the selection or arrangement independently (i.e., without copying that selection or arrangement from another work), and that it display some minimal level of creativity*" (U.S. Supreme Court, 1991). Although copyright law in the United States differs from systems like droit d'auteur in Civil Law countries (Benhamou and Farchy, 2009), the originality analysis used here can also inform assessments in other jurisdictions, such as Brazil.

Section 7 of the Brazilian Copyright Act provides an illustrative list of elements deemed eligible for protection. This serves as a reference point for interpreting the originality criterion, as the legislator presumes that any element included in this list is original and must be protected. For example, literary, artistic or scientific works may be considered authorial works (Section 7, I), musical compositions with or without lyrics (Section 7, V), and works that, by their selection, organization, or arrangement, constitute an intellectual creation (Section 7, XIII).

2.2 Applicability of the originality criteria to *prompts* and the limits imposed by the exclusion of ideas, methods, and functionality

After outlining the originality criterion, this paper examines whether a prompt satisfies the requirements for protection under the Brazilian Copyright Act. The concept of originality is interpreted broadly and subjectively. Therefore, evaluating prompts' eligibility for copyright protection requires an individual analysis of each prompt.

If the prompt contains original content, it may be protected by copyright. This is because a prompt may contain elements that demonstrate the author's creativity, differentiating it from other existing prompts. For example, the *prompt* below contains several elements of the author's subjectivity, which are associated with his experience, style and experience, and which would represent an independent and creative creation (the "creations of the spirit" of Section 7 of the Brazilian Copyright Act , with the requirements defined by the STJ and the Supreme Court of the United States in the cases mentioned above):

Imagine a futuristic city built in harmony with nature, where glass skyscrapers intertwine with gigantic trees and hanging gardens. The city is illuminated by bioluminescence, with rivers of light flowing through the streets and crystal bridges connecting different neighborhoods. In the center of the city, there is a square where artists from all galaxies gather to display their works, and the air is always filled with a symphony of natural and technological sounds. Describe the daily life of the inhabitants of this city, their interactions, festivals, and how technology and nature coexist in perfect harmony.

However, there would be arguments to refute this protection. Prompts (including more complex ones, such as the example above) are, at their core, textual instructions addressed to a model. Similarly, these instructions may be handled like ideas shared by an employer with an employee for creating a musical arrangement, or by a client commissioning work from an artist. In these cases, under the Brazilian Copyright Act, it could be argued that what is subject to protection is only the result (which would be comparable to AI-generated content), whose authorship and ownership would, by *default, be* attributed to the employee or contractor, unless, in relation exclusively to ownership, there is a written provision to the contrary.

The Brazilian Copyright Act states that "ideas" are not eligible for copyright protection (Section 8, I). To be protected, ideas must be expressed and fixed in tangible or intangible media, at which point a protected work comes into existence. Regarding the prompt above, one could argue that the instructions it includes would serve only as suggestions for generating textual content through generative AI.

The reluctance to recognize prompts can also be based on the merger doctrine. Merger doctrine is often invoked in the U.S. to avoid de facto monopolies over ideas when there are only limited ways to express them. While Brazilian law does not label this doctrine per se, the same policy appears in the Brazilian Copyright Act, as mentioned above.

When an idea can only be expressed in very limited ways, copyright protection of the expression would, in effect, amount to protection of the idea itself, which is prohibited both by Brazilian legislation and by international treaties. In the case of AI prompts, the application of the merger doctrine means that simple, direct, or functional instructions - whose formulation is virtually inevitable - should not be

protected by copyright. The U.S. Copyright Office recently invoked this reasoning to deny protection to prompts, framing them as mere functional instructions.[106]

The assertion that every *prompt* constitutes an idea can be challenged. *Prompts* can be just ideas, but not in every case. A sophisticated *prompt*, such as the example above, could be considered more than just an idea due to its level of detail and creativity. The originality and the specific way in which the *prompt* was prepared should be able to give it copyright protection, as it would represent a unique and creative expression of the author, almost impossible (unless by enormous coincidence) to be replicated, if not from a copy. The important distinction that is made here is between the general idea (in the case of the example, a futuristic city in harmony with nature) and the specific expression of that idea (in the case of the example, the detailed and creative description of the city, its elements, and the daily life of the inhabitants).

The US Copyright Office's position suggests that the unpredictability of AI outputs, even when identical prompts are used, demonstrates a lack of human control and, by implication, undermines the originality or protectability of the prompt itself. Rather than a conflation, U.S. practice tends to stress the unpredictability of the output as a reason to deny protection, whereas Brazilian analysis grounded in *droit d'auteur* should focus on the originality of the input. Both pints can coexist; the comparative contrast simply highlights different emphases. The fact that a prompt can generate a wide range of outputs does not diminish the creative effort or originality embedded in the prompt's formulation. To draw a parallel, film scripts are widely recognized as protectable works under copyright law, even though screenwriters have no control over how different directors, actors, or production teams might interpret and realize the script. The script's protectability is not contingent on the uniformity or predictability of its interpretation, but rather on the originality and creative expression of the author. Similarly, a prompt may serve as a creative blueprint, and its lack of determinism over the final output does not negate its status as an original work deserving of protection. The essence of copyright lies in safeguarding the author's unique expression, not in guaranteeing a fixed or foreseeable result.

Still, certain *prompts* will lack originality and/or may represent simple ideas. These are the cases of everyday questions (a *prompt* that, for example, asks the generative AI system to perform a certain mathematical calculation or that lists the capitals of all the countries on a given continent) or *prompts* that contain general ideas (such as the one that, still in the context of the highlighted example, asks generative AI simply to "imagine a futuristic city in harmony with nature and write about it"). In these situations, the *prompts* will not deserve the copyright protection provided for in the Brazilian Copyright Act.

It should be noted that the possible copyright protection of *prompts* defended in this paper should not be limited only to cases involving instructions commonly associated with the artistic or literary world. The Brazilian Copyright Act also protects works that by their selection, organization, or arrangement constitute an intellectual creation (Section 7, XIII). This can also be the case with *prompts*. A *prompt* that contains various elements that individually are not subject to protection (such as facts, figures, etc.), but that has been organized originally, also deserves the protection of the Brazilian Copyright Act. This would be the case, for example, with extremely detailed prompts used to perform professional tasks (for example, to create marketing presentations or generate financial reports with particular characteristics).

To defend this argument, one can use, by analogy, the interpretation of the aforementioned *Feist case* of the Supreme Court of the United States, which examines, in general terms, whether or not telephone contact lists would be subject to copyright protection. Although it was concluded that the lists examined in the case were not protectable, the Supreme Court clarified that: *"[f]actual compilations, on the other hand, may possess the requisite originality. The compilation author typically chooses which facts to include, in what order to place them, and how to arrange the collected data so that they may be used effectively by readers. These choices as to selection and arrangement, so long as they are made independently by the compiler and entail a minimal degree of creativity, are sufficiently original that Congress may protect such compilations through the copyright laws"* (U.S. Supreme Court, 1991).

In the same vein, the STJ also recognized the originality of dictionaries in the light of the Brazilian Copyright Act, due to the way in which the entries and all their content are organized: *In other words, all human knowledge derives from discoveries that preceded it, and whose content can be used in later works. That is not what originality consists of. Not every work of creation of the human spirit is totally original, in its strictest concept, but originality stems from the way in which its content is exposed. This*

[106] According to the U.S. Copyright Office "Prompts essentially function as instructions that convey unprotectible ideas." Available at: https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf. Aug, 2025

is precisely the case of the dictionary: although the concepts could be known, the organization of this content in entries, in the way it was made, is what demonstrates the originality of the work (Superior Tribunal de Justiça, 2019).

In these cases, however, the scope of protection conferred by copyright is quite limited, since the elements of the content of the work are not protected. This conclusion is a consequence of a lower degree of originality, relative to works composed entirely of elements that in themselves denote creativity (such as the *prompt* in the example at the beginning of this Section). This is also a teaching given in the Feist case, when clarifying the fact-expression dichotomy in authorial works that contain facts: *Since facts do not owe their origin to an act of authorship, they are not original, and thus are not copyrightable. Although a compilation of facts may possess the requisite originality because the author typically chooses which facts to include, in what order to place them, and how to arrange the data so that readers may use them effectively, copyright protection extends only to those components of the work that are original to the author, not to the facts themselves. This fact/expression dichotomy severely limits the scope of protection in fact-based works* (U.S. Supreme Court, 1991)/

Some authors may challenge the copyright protection of prompts - whether artistic works or original compilations - from a stricter perspective, arguing that copyright should not cover works with any functional aspect. For example, for Barbosa (2013) copyright should not restrict any functional equivalent works. Thus, literary, artistic, or scientific works should not have any functionality beyond their purpose of expression, but should be produced with the sole purpose of expressing ideas, concepts, and sensations. This understanding was adopted by the STJ, which, when examining the possibility of recognition of copyright over legal professional documents (such as court briefs), understood that procedural documents are utilitarian and that, therefore, they are not protected by the Brazilian Copyright Act (Superior Tribunal de Justiça, 2002). Considering that *prompts* always *have*, by their nature, a utilitarian character [107] (*i.e.,* that of instructing a generative AI system to generate content), from this point of view, no *prompt* would deserve the protection provided by the Brazilian Copyright Act.

This paper presents an alternative perspective to this restrictive view, considering that it is inconsistent with the relevant legislation. The Brazilian Copyright Act does not exclude from its list of protection any and all works that have a functional character, but only those provided for in the list of exceptions to copyright protection in Section 8 (such as game rules and methods). On the contrary, item VII of the aforementioned Section 8 allows an interpretation, *a contrario sensu,* that the legislator recognized the possibility that an authorial work exercises some functionality. This item determines that the industrial or commercial use of ideas contained in the works is not subject to copyright protection. Thus, a work can provide ideas to others about ways to perform a certain activity – what is functional and utilitarian. To illustrate, a cookbook is an authorial work protected by the Brazilian Copyright Act, but no one can prohibit a confectioner from selling a candy made from a recipe from this book on the grounds of violation of the copyright of its author (or the holder of the respective patrimonial rights). Furthermore, the Brazilian Copyright Act recognizes that computer programs are within its scope of protection (Section 7, XII), while the Software Law, when defining them, leaves no room for doubt about their functionality: *Section 1 A computer program is the expression of an organized set of instructions in natural or coded language, contained in a physical support of any nature, of necessary use in automatic information processing machines, devices, instruments or peripheral equipment, based on digital or analogous technique,* ***to make them work in a certain way and for specific purposes****. (emphasis added)*

Therefore, prompts, even if functional, can be protected by the Brazilian Copyright Act, regardless of whether it is a literary work or a compilation of elements organized by the person who wrote it, as long as its final content is original, under the terms discussed throughout this Section 1.

3 Introductory Discussion on Legal and Social Consequences of Prompt Protection

A comprehensive analysis of prompt authorship must go beyond doctrinal questions and consider the broader legal and social ramifications. The recognition or denial of prompt authorship has the potential to reshape regulatory frameworks, user rights, trust in digital systems, and the governance of digital authorship. This Section 2 explores these dimensions, drawing connections to the arguments developed

[107] And here, a recall of the definition proposed by the article for *prompts*: a *prompt* is a textual content used by the user of generative AI to have generative AI generate new content.

in the first part of this paper and incorporating practical examples and challenges identified in the previous discussion.

3.1 Legal Effects: Regulation and User Rights

3.1.1 Challenges to Regulation and Enforcement of Copyright Law

Recognizing prompt authorship under copyright law would require adaptations in existing regulatory and enforcement frameworks. Legislators, regulators, and the Judiciary would need to clarify the boundaries of protection, especially to avoid the monopolization of ideas or functional instructions, which are excluded from copyright protection. The challenge lies in distinguishing between protectable creative expressions and unprotectable ideas or methods, a distinction that is not always clear-cut in the context of prompts.

The recognition of prompt authorship could prompt the development of new licensing models, contractual arrangements, and dispute resolution mechanisms. For instance, prompt marketplaces may need to implement robust systems for tracking ownership, licensing terms, and permissible uses, while also ensuring compliance with data protection and competition laws. The risk of overlapping claims - such as when multiple users independently create similar prompts - would require careful attention.

Furthermore, the recognition of prompt authorship may have downstream effects on the protection of AI-generated outputs. If the copyright on a particular prompt is assigned to its creator, only that person can exploit it, and it is only from this exploitation that content can be generated by generative AI systems. By holding not only the authorship, but also the prerogatives to authorize the use of the prompt and define the conditions of such use, the author will also have some control over the content to be generated (and may provide, for example, in licensing agreements for their prompts, certain rights and limitations in relation to the content generated by them). The possibility of recognizing the copyright protection of certain prompts demystifies the maxim that in Brazil, currently, there is no legal means capable of protecting content generated by generative AI. By assigning protection to a prompt, the author is also given control over the content that it is able to generate.

3.1.2 User Rights

The legal status of prompts directly impacts user rights. If prompts are recognized as protected works, users may gain enforceable rights to attribution ("moral rights" under Civil Law), remuneration, and control over the use and dissemination of their prompts (property rights). This could empower users, granting them greater agency in the digital ecosystem and incentivizing the creation of high-quality, innovative prompts. [108] [109]

The practical enforcement of these rights presents significant challenges. The bilateral and often private nature of prompt submission to AI systems complicates the detection and proof of infringement. Additionally, the widespread practice of using prompts to train AI models raises questions about the scope of user consent and the potential for unauthorized exploitation of user-generated content. For example, many generative AI platforms process prompts not only to return results to the user but also to train their models and generate new results for other users.[110] This creates a scenario in which a user who has licensed a prompt for limited use may find it repurposed or disseminated beyond the agreed terms, especially if the platform's terms of service allow for such use.

Some companies, such as OpenAI, already provide opt-out mechanisms, allowing users to indicate that they do not want their content (including prompts) used for training purposes.[111]However, the effectiveness of such mechanisms in protecting user rights remains to be seen, especially given the

Another challenge concerns the difficulty of detecting and tracing the use of prompts for purposes such as demonstrating copyright infringement. Since interactions with prompts typically occur only

[108] Prompt engineering, *which consists of the discipline to develop and optimize* prompts *so that they return increasingly satisfactory results for their users, has already become popular and can be used to develop prompts* with economic value. See, for example, https://www.*promptingguide.ai/pt* available Jan. 2025. There are already several *marketplaces* specializing in the sale of *prompts*. See, for example, VibePrompts and *PromptBase*.

[109] Some jurisdictions fear that protection prompts could chill downstream innovation or impede interoperability by extending exclusive rights to functional instructions. This paper takes that concern seriously and proposes a narrow, "thin" or even non-applicable protection for functional or highly constrained prompts, while reserving fuller protection for genuinely expressive prompts

[110] For example, OpenAi's Terms of Use state that: "Our Use of Content. We may use your Content to provide, maintain, develop, and improve our Services, comply with applicable law, our terms and policies, and keep our Services secure." (emphasis added) Available at https://openai.com/pt-BR/policies/terms-of-use/ Jan. 2025.

[111] "Opt out. If you do not want us to use your Content to train our models, you can opt out by following the instructions in this Help Center article (opens in a new window). Please note that in certain cases this may limit the ability of our Services to respond more appropriately to your specific needs." (emphasis added) Available at https://openai.com/pt-BR/policies/terms-of-use/ Jan. 2025.

between the user and the AI system, any violation will rarely be public. This raises questions about how the copyright owner can identify and prove infringement, and whether new technical or legal mechanisms are needed to facilitate enforcement.

3.2 Social Effects: Trust, Inclusion, and Digital Authorship Governance

3.2.1 Trust in AI Systems.

Clarity regarding the authorship and protection of prompts can significantly influence user trust in generative AI platforms. Studies on human-AI interaction (Sivaraman, Bukowski, Levin, Kahn, and Perer, 2023; Souverain, 2025) demonstrate that users are more likely to engage meaningfully with AI systems when their contributions are recognized and respected. Conversely, ambiguity or lack of protection may deter users from sharing valuable prompts, stifling creativity and innovation.

Transparent governance structures that recognize and respect user authorship can foster a sense of fairness and accountability, which are essential for building sustainable digital ecosystems. This is particularly relevant in contexts where users contribute to the training and improvement of AI models, often without explicit recognition or compensation.

3.2.2 Inclusion and Social Value

The attribution (or lack thereof) of prompt authorship has broader social implications. Recognizing the creative labor involved in crafting complex prompts can validate the contributions of diverse user groups, promoting inclusion and democratizing access to digital authorship. It can also encourage participation from underrepresented communities, who may otherwise be excluded from the benefits of AI-driven creativity.

On the other hand, failure to acknowledge prompt authorship risks perpetuating existing inequalities, as the value generated by users may be appropriated by platform operators or developers without adequate recognition or reward.

3.2.3 Governance of Digital Authorship

The definition and management of prompt authorship are central to the governance of digital authorship in AI-mediated environments. Clear rules regarding ownership, attribution, and exploitation of prompts can enhance transparency, facilitate the resolution of disputes, and support the development of fair and inclusive digital markets.

At the same time, the governance of prompt authorship must balance the interests of individual creators, platform operators, and the broader public. Overly restrictive protection could hinder the free flow of ideas and impede collaborative innovation, while insufficient protection may undermine user trust and participation.

4 Conclusion

Section 1 of this paper has demonstrated that there is clear scope within the current Brazilian Copyright Act for the recognition of copyright protection for prompts, provided they meet the criterion of originality. The analysis undertaken reveals that, while not all prompts will qualify as protectable works, those that embody creative expression, narrative vision, or distinctive organization may indeed fall within the ambit of copyright protection as defined by Brazilian law. This recognition, however, is not merely a technical legal matter; it gives rise to a host of legal and social questions that merit further scrutiny.

Having achieved the primary aim of setting up the legal plausibility of prompt protection in Brazil, this study has also outlined, in Section 2, a range of scenarios and consequences that flow from such recognition. These include challenges related to the enforcement of rights, the delineation of ownership, the impact on user agency, and broader questions of governance and trust in digital environments. These issues are complex and multifaceted, intersecting with ongoing debates about the regulation of artificial intelligence, the evolution of digital authorship, and the inclusivity of creative ecosystems.

Accordingly, the scenarios and questions raised in this paper should be viewed as starting points for deeper academic inquiry. Further research is essential to clarify the boundaries of protection, develop effective enforcement mechanisms, and assess the broader societal implications of recognizing prompt authorship. By situating the discussion within the existing legal framework and highlighting areas for future exploration, this paper aims to contribute to a more nuanced and forward-looking understanding of digital authorship in the age of generative AI.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Barbosa DB (2013) Copyright. Lumen Juris, Rio de Janeiro

Benhamou F, Farchy J (2009) Droit d'auteur et copyright. La Découverte, Paris

Bittar CA (1992) Copyright

Lee E (2023) Prompting progress: Authorship in the age of ai. SSRN Electronic Journal

Mazzi F (2024) Authorship in artificial intelligence-generated works: Exploring originality in text prompts and artificial intelligence outputs through philosophical foundations of copyright and collage protection. The Journal of World Intellectual Property

Paranaguá P, Branco S (2009) Copyright. FGV, Rio de Janeiro

Schirru L (2020) Copyright and artificial intelligence: authorship and ownership in ai products. PhD Thesis, Institute of Economics (UFRJ), Rio de Janeiro

Sivaraman V, Bukowski LA, Levin J, et al (2023) Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, pp 1–18, 10.1145/3544548.3581075, article 754

Souverain T (2025) Ai to renew public employment services? explanation and trust of domain experts. AI Ethics 5:55–70. 10.1007/s43681-024-00629-w

Rocha de Souza A, Schirru L, Lana AP, et al (2024) Artificial intelligence and copyright: Contributions to the regulatory debate in brazil

Superior Tribunal de Justiça (2002) REsp 351358/DF, Rel. Min. Ruy Rosado de Aguiar, 4ª Turma. Julg. 04/06/2002; Publ. 16/09/2002

Superior Tribunal de Justiça (2019) REsp 1756242/SP, Rel. Min. Ricardo Villas Bôas Cueva, 3ª Turma. Julg. 19/11/2019; Publ. 28/11/2019

Superior Tribunal de Justiça (2020) REsp 1704189/RJ, Rel. Min. Ricardo Villas Bôas Cueva, 3ª Turma. Julg. 13/10/2020; Publ. 19/10/2020

Tribunal de Justiça de São Paulo (2014) Apelação Cível 1022055-65.2013.8.26.0100, Rel. Des. Mendes Pereira, 7ª Câmara de Direito Privado. Julg. 14/05/2014; Publ. 20/05/2014

Tribunal de Justiça do Rio Grande do Sul (2017) Apelação 70073070914, Rel. Des. Isabel Dias Almeida, 5ª Câmara Cível. Julg. 26/04/2017; Publ. 02/05/2017

U.S. Supreme Court (1991) Feist Publications, Inc. v. Rural Telephone Service Co. 499 U.S. 340

Critical Thinking and Media Literacy in Youth from Underserved Contexts: Evidence from a Field Intervention

Alice Bougnères[1], Hugo Besançon[1], Melisa Basol[2]

[1] Square, Paris, France. [2] Pulse, London, United Kingdom.

Corresponding author. E-mail: alice@squaremail.org

Abstract

Misinformation poses a threat to democratic resilience and informed citizenship, yet most interventions and studies focus on highly educated or already-engaged populations. This paper examines the work carried out with a sample of 630 young adults from disadvantaged backgrounds in Île-de-France, France. It offers unique insights into gaps in media literacy, critical thinking, civic confidence, and misinformation detection, underscoring the urgent need for multi-layered and evidence-based initiatives, while also assessing the impact of political and media literacy workshops.

The data highlight a significant trust deficit with only 9.5% of young adults reporting trust in institutions, further exacerbated by endorsement of conspiracy beliefs, with almost 40% agreeing with conspiratorial statements. Square offered media literacy workshops spanning 8 to 20 hours over 1.5 to 4 days to young adults in 2023 and 2024. The trainings aimed to develop critical thinking skills, support the ability to analyse information, and foster civic engagement of the participants.

They demonstrated clear improvements in defining key media literacy concepts. 79.4% reported checking sources more frequently and 73.8% said they would be seeking out diverse perspectives, suggesting an increased awareness of verification practices. They translated intention into action, maintaining their commitment to critical engagement with online content for up to three months. The report reinforces that simply raising awareness is not enough and that deep-rooted vulnerabilities cannot be addressed through short-term interventions alone. Still, the findings illustrate that short-term media literacy interventions can effectively raise awareness and enhance conceptual understanding critical to navigating digital environments.

Keywords

Media Literacy, Critical Thinking, Civic Confidence, Political Efficacy, Credibility Assessment, Misinformation Detection.

1 Introduction

This paper focuses on young adults' critical thinking abilities before and after taking part in Square's media literacy workshops. Between 2023 and 2025, Square partnered with local NGOs and employment organisations to support young adults from disadvantaged neighbourhoods in Clichy-sous-Bois and Mantes-la-Jolie, France, areas marked by high poverty and unemployment. Despite numerous media literacy and critical thinking interventions, comprehensive evaluation work and research of their effectiveness remain scarce. This is in part due to the lack of a standardised definition for media literacy and critical thinking (Potter, 2022), and to the variety of actors involved (NGOs, policymakers, academia, formal and informal education providers, media organisations, supra-national entities…). The wide array of intervention formats and target audiences further complicates robust evaluation efforts (Edwards et al., 2025).

Square's work aims to fill this gap and provide an overview of the real habits of young people from disadvantaged areas with a focus on media literacy gaps. The baseline data highlight profound distrust in institutions, endorsement of conspiracy beliefs, and motivated reasoning, where many reject information that contradicts their views. Combined with uncertainty in evaluating online content and weak verification habits, these challenges make a clear case for structured media literacy education. These insights establish a foundation for assessing the long-term impact of our workshops. We then looked at the development of critical thinking by analysing essential skills post-workshop such as knowledge of key concepts, real ability to differentiate facts from opinions and positive changes in behaviours.

The results come from over 3 years of interventions aimed at developing critical thinking skills, supporting the ability to analyse information and fostering civic engagement of the participants. Square's approach combines 'prebunking' methods (where participants are put in a situation where they are confronted with manipulation techniques to learn to identify and resist them) with metacognition. Participants were shown various types of manipulative content and questioned on its meaning, the reasons why it would be posted and shared, as well as helped in identifying the technique used to influence from both an emotional and technical standpoint. Key topics included online manipulation techniques, freedom of speech, verification tools, distinguishing facts from opinions, understanding the context surrounding common manipulative content (false contexts, deepfakes, anecdotal proofs…) and discussing biases. Content and examples used were constantly updated throughout the length of the project to best fit the interests and realities of the target audience and what they see online. Depending on context, Square offered workshops ranging from 8 to 20 hours over 1,5 to 4 days for young adults.

2 Method

We aimed to test whether short "inoculation" (Roozenbeek et al., 2022) interventions using "prebunking" techniques could produce measurable and sustainable impact in resisting manipulative content online and in trusting authentic information. Much like a vaccine, the idea is to expose participants to common misinformation tactics and help them become more resilient to manipulative content by developing critical thinking skills.

This study also examines various psychological and civic dimensions, including self-esteem, open-mindedness, political efficacy, conspiracy thinking, and institutional trust. The aim was to determine their potential influence on participants' critical thinking skills, information-seeking behaviors, and resilience to online manipulation and disinformation. This approach allowed for an investigation into how fundamental civic, social, and political factors shape online interactions and information access within the cohort.

We also looked at participants' veracity assessment skills (Guay et al., 2023). Current research focuses on using a "discernment measure" by exposing participants to true and false statements and having them assess the reliability of those statements. We then differentiate between the perceived reliability of the true and false statements; the discernment score being positive when true statements are perceived as more reliable than false ones. In line with recent work (Pfänder and Altay, 2025) we wanted to increase the accuracy of these measures by using ecological content tailored to our audiences in a format resembling what they see every day on social media. We used screenshots of social media posts and a Likert scale consisting of a statement followed by a range of response options ranging from a negative extreme to a positive extreme to measure the perceived reliability of the content, while leaving open the comment sections for participants to explain their choices. Uncommon in the literature, this option

allowed for more qualitative analysis and understanding of the reception and comprehension of the content as well as the choices made by the participants.

To evaluate Square's workshop's impact on the perceptions and knowledge of young people, we conducted pre- and post-workshop questionnaires[112]. The pre-workshop questionnaire served a dual purpose: it was designed to gather data on a number of attitudes and behaviours towards social media and it aimed to establish a baseline level on a number of media literacy dimensions which were then compared in the post-workshop questionnaire.

The total baseline sample reached up to 630 young adults, though the exact number of answers varies across different metrics. Some items were introduced, removed, optimised, or adjusted throughout data collection, and participants could choose not to answer certain questions, leading to slight variations in sample sizes for specific measures. The number of responses will be specified throughout the paper for each dimension "(N = x)"[113].

The end-of-workshop questionnaire mirrored the one given at the start, covering the same dimensions of interest. This allows for a before-and-after comparison for each dimension. This method is based on the assumption that young people's perceptions and knowledge would have remained the same if there had been no workshops. As the time between the two questionnaires is short, 4 days maximum, this reduces the probability that events will significantly alter young people's perceptions and knowledge outside the workshops.

We also collected follow-up data 6 to 12 weeks after the workshops, which we present here for completeness, but attrition is very high, which reduces the scope of the analyses. It is important to note that some evaluation items changed during the project, one of the aims of which is precisely to test evaluation items, complicating comparisons across testing waves.

The following sections present insights into young adults' knowledge retention, reasoning abilities, and behavioural shifts.

3 Results

Our 630 young adults' sample is balanced in terms of gender, with 53% male and 47% female. The age distribution ranges mostly from 18 to 25 years old. Most participants hold a pre-high school diploma - Brevet des collèges (22.2%), followed by high school diplomas - Bac Pro (13.3%) and Bac General or Technologique (11.9%), with higher education levels (2 years post-graduate) being less represented at 8.7%.

3.1 Media literacy gaps in young adults

The first part of this paper focuses on the pre-workshop data, showcasing young participants' levels of trust towards institutions, sense of agency, civic confidence, verification habits, open-mindedness, media literacy knowledge, conspiratorial thinking and ability to discern veracity across different formats.

3.1.1 Profound institutional distrust (N = 477)

Our results reveal a profound institutional trust gap among our cohort compared to the national average. A significant trust deficit is evident with young adults reporting significantly lower trust in institutions (9.5%) than the national average (CEVIPOF, 2024) in France. Critically, the least trusted institution is political parties, reflecting strong disillusionment with formal governance. The starkest disparities appear in law enforcement (7.9% vs. 70%) and the judicial system (7.2% vs. 45%), suggesting that our cohort may feel particularly alienated from the systems tasked with upholding justice and security. While the most trusted messengers among young adults are researchers and scientists (22.7%) and schools (15.8%), even these figures fall significantly below the national averages (81% and 67%, respectively).

Our analyses suggest a strong negative correlation between government trust and conspiracy beliefs, meaning that lower trust in government is associated with higher endorsement of conspiratorial thinking.

[112] For more details, please refer to the URL where the complete questionnaires are stored: https://osf.io/bfpd2/?view_only=5e1d0e38984247f79a618b8e97992bbc

[113] For more details, please refer to the URL where the complete tables are stored: https://osf.io/bfpd2/?view_only=5e1d0e38984247f79a618b8e97992bbc

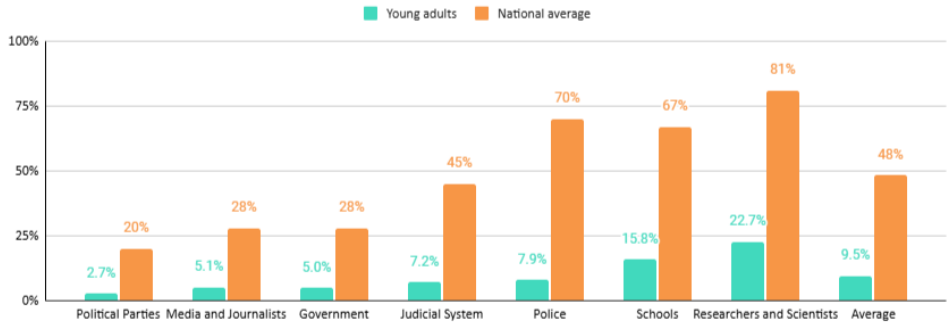


Fig. C1: A significant lack of declared trust in institutions among the survey population, compared to the French national population

3.1.2 Moderate confidence in the sense of agency (N = 467)

Our cohort demonstrates moderate confidence in both personal and collective agency, though their belief in the ability to change society through action is notably lower than the national trend. Their confidence in personal autonomy is closer to the national average (CEVIPOF, 2021), suggesting they feel more control over their own lives than over broader societal change. We must however consider that these national averages often focus on the general population (and are thus not limited to specific age ranges), making direct comparisons difficult. Still, this gap in collective agency may reflect socioeconomic barriers, institutional distrust, or limited civic engagement opportunities, which could impact their willingness to participate in public life.

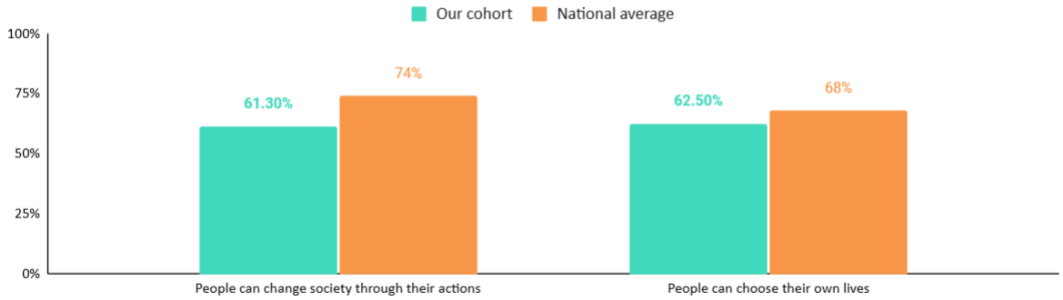


Fig. C2: Limited confidence in participants' sense of collective and personal agency compared to the national average

3.1.3 Lack of civic confidence (N = 467)

Young adults display a worrying lack of civic confidence (Schulz et al., 2010) and consistently report feeling less knowledgeable and less engaged in political discussions compared to the national average (CNESCO, 2018). A particularly stark gap appears in their ability to express political opinions and participate in debates, suggesting that many feel unprepared or excluded from political discourse. Furthermore, young adults say they struggle with understanding political questions at a much higher rate than the general population, which may contribute to their disengagement. This trend is consistent with data on middle schoolers, where civic trust is already low, and here too, we witness a further decline in confidence and engagement as young people grow older. These insights point to a broader issue of political alienation, underscoring the need for targeted efforts to enhance political education and engagement among young people.

3.1.4 Not looking for trustworthy content (N = 469)

Whereas 58.9% young adults report low or no confidence in the information they see on social media, it is still where they mainly get their information from, suggesting that they are not looking for the truth. They evolve in an informational space they seem to knowingly not trust, hinting that they might not be looking for trustworthy information. Despite the scepticism towards information quality on these

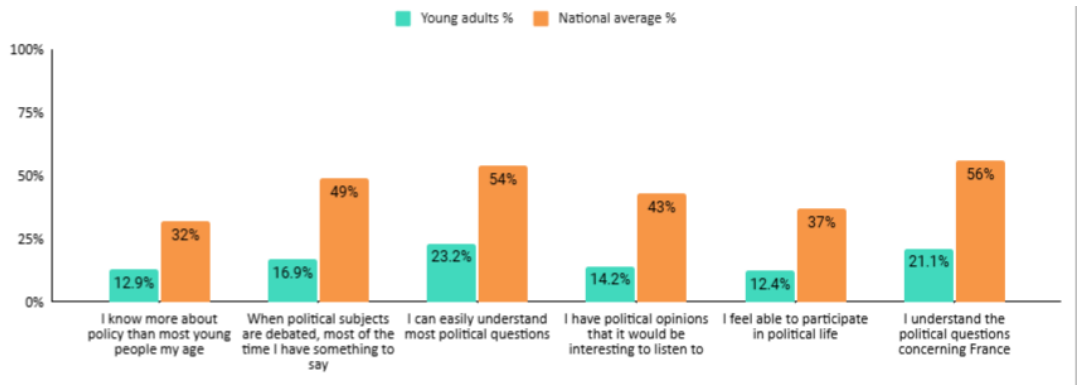


Fig. C3: Lack of civic confidence in Square's workshop participants' compared to the national average

platforms, users continue to engage, driven by a deep-seated desire for social connectivity and identity affirmation (Nera, 2023).

Furthermore, we found limited critical engagement with information. Only 26.3% actively check the sources of the information they read, and just 24.5% report engaging with viewpoints different from their own, suggesting a risk of echo chambers and passive consumption of information.

There are gaps in verification habits: our cohort displays a heavy reliance on Google and general internet searches, while fact-checking services and traditional media remain underutilised. Few cite official sources, and some turn to non-traditional platforms like ChatGPT and TikTok, reflecting a lack of structured verification methods (N = 78).

3.1.5 Tension between conviction and adaptability (N = 464)

Open-mindedness, as quantified by the Actively Open-minded Thinking (AOT) scale (Stanovich and Toplak, 2023), is argued to be essential in decision-making, belief evaluation, and evidence assessment. This psychological construct encourages the examination of different viewpoints, challenges personal beliefs, and adapts thinking based on new information, which is vital for sifting through complex information environments. Research has demonstrated that open-mindedness is associated with improved decision-making and reduced cognitive biases, fostering a more nuanced processing of new information and reducing biases, judgment errors, and dogmatism (Haran et al., 2013) (De Keersmaecker and Roets, 2017) (Pennycook et al., 2020).

The data reveals a tension between openness to new information and resistance to change. While many value reconsidering conclusions and exploring multiple perspectives, a significant portion prioritise conviction over adaptability, with 45.8% believing in holding onto beliefs despite contradicting evidence and 25.2% finding it acceptable to ignore counterevidence. Encouragingly, doubt is widely accepted, but the persistence of views favouring confidence over questioning emphasises the need to strengthen critical thinking skills.

3.1.6 Limited knowledge of the information landscape (N = 253)

While almost 70% of workshop participants do not know what an algorithmic bubble is, less than 50% of the cohort correctly defined terms such as disinformation, misinformation, and deepfake before the workshop. Results show significant uncertainty, particularly regarding platform regulations and algorithmic influence. Only 30.9% correctly identified that platforms are legally required to participate in the fight against disinformation, while 58.3% expressed uncertainty. Similarly, while 51.6% correctly recognised that algorithms filter the information displayed on social media, 37.3% were unsure.

These findings underscore a lack of foundational knowledge about digital governance and suggest a need for more comprehensive digital literacy education that explicitly integrates legal rights, platform governance, and algorithmic influence.

3.1.7 Difficulties to navigate between facts and opinions (N = 314)

The ability to effectively distinguish between statements of fact and expressions of opinion is a key component of critical thinking and media literacy. Participants were presented with a range of statements, spanning various topics from sports and personal preferences to social and scientific assertions, to assess their ability to make the distinction.

Statement	% Agreement
Real experts are prepared to admit that they are not sure they know the truth	40.60%
People should consider the evidence against the conclusion they prefer	42.80%
Doubting is the result of confused thinking	36.30%
People should reconsider their conclusions when new relevant information becomes available	56.50%
Changing your mind is a sign of weakness	8%
People need to actively look for reasons why they might be wrong	50.60%
It is acceptable to ignore evidence that contradicts your beliefs	25.20%
It is important to be faithful to your beliefs even when evidence is brought against them	45.80%
There's nothing wrong with doubt on many issues	61%
When faced with a difficult question, try to think of several possible answers before coming to a conclusion	67.40%
It is better to be confident in a conclusion even when you have good reason to question it	40.50%

Fig. C4: Table showcasing the level of participants' open-mindedness

While straightforward, widely known facts, such as the number of letters in the alphabet, were more likely to be correctly identified, responses indicate greater uncertainty in areas that require deeper contextual knowledge.

For example, only 48.5% correctly stated that "water boils at 100 degrees" is a fact, while the lowest correct identification rate was for the statement that "France is a secular republic and, within this framework, the law does not recognise blasphemy and offence", with just 25.1% answering correctly and 40.7% uncertain. While participants show confidence in widely accepted factual knowledge, they are noticeably more cautious when assessing statements related to geopolitics, governance or legal principles. This uncertainty may reflect intellectual caution and, at times, identity-based reasoning, rather than outright misinformation, suggesting an awareness of their own knowledge limitations.

3.1.8 Conspiratorial thinking rooted in civic disengagement and institutional distrust(N = 459)

We used the Adolescent Conspiracy Beliefs Questionnaire (ACBQ) (Jolley et al., 2021), aimed at measuring adolescents' beliefs in conspiracy theories. Data reveal that mistrust runs deep, with almost 40% of our cohort agree with conspiratorial beliefs, with many believing that authorities withhold information, engage in surveillance, and allow harmful events to unfold deliberately. Public health conspiracies, such as governments weaponizing diseases, also see notable support. While extreme theories, such as 5G microchips in vaccines, remain less common, scepticism towards major political and historical events, including election fraud and state-orchestrated attacks, suggests a broader pattern of institutional distrust.

These results highlight significant scepticism toward institutional motives, which aligns with broader trends of low institutional trust and limited civic confidence within the cohort. Conspiratorial thinking has far-reaching consequences, shaping individuals' perceptions of reality, influencing decision-making, and undermining trust in public institutions, health initiatives, and democratic processes (Kovic and Füchslin, 2018). Among young people, these beliefs can reinforce civic disengagement, scepticism toward credible sources, and increased susceptibility to misinformation.

Statement	% Agreement
The government is deliberately hiding important information from the public.	69.50%
The government is secretly monitoring people.	55.70%
Some political groups have secret plans that are not good for society.	63.70%
Some diseases have been created by the government to be used as weapons.	40.70%
The government often knows about terrorist attacks and lets them happen.	29.60%
Governments have deliberately allowed diseases to spread to certain groups of people.	40.50%
Secret groups control people's minds without their knowledge.	30%
Secret societies control politicians and other leaders.	40.50%
Secret societies influence many political decisions.	49.90%
The Covid-19 vaccine contains 5G microchips to track us.	10.60%
The American government lied to justify the war in Iraq.	42.10%
It is certain that the Charlie Hebdo attacks were planned and carried out by the government.	50.30%
Experts talk about climate change to limit economic development.	24.60%
The result of the 2020 U.S. presidential election was rigged from the start.	23.40%
A vast network of corruption and money laundering exists within the government.	17.30%

Fig. C5: Table showcasing participants' answers when faced with conspiratorial statements. This table includes all conspiratorial statements tested over recent years. The first 9 statements are from the ACBQ and have up to 450 responses. The last six statements were discontinued, so sample sizes here are up to 233.

3.1.9 More sceptical than gullible

To complement our assessment of perceived online literacy skills, we evaluated participants' ability to assess veracity using a series of psychometrically validated misinformation items and real-world social media content. The stimuli ranged from text-based headlines and statements to more complex social media screenshots, incorporating engagement metrics, comments, and source information. This approach allowed us to assess not only their ability to evaluate isolated factual claims but also their susceptibility to misinformation in a more dynamic, digital environment that mirrors their everyday online interactions.

We presented them (N = 187) with a series of statements drawn from the Misinformation Susceptibility Test (MIST). This assessment applies a misinformation susceptibility test (MIST) scale, adapted to include locally relevant claims reflecting the political, media, and scientific narratives shaping young adults' information environments. A critical departure from the original scales is that we introduced an "I don't know" response option, which dominates across multiple items. However, some research suggests that the inclusion of "I don't know" options can negatively impact data quality, potentially leading to less valid responses (Krosnick et al., 2002). Despite this, we believe it was essential to include this option here, as it allowed participants to express uncertainty rather than being forced into a false dichotomy,

leading to a more nuanced understanding of their knowledge gaps. To our knowledge, this is the first application of such an option in MIST, aiming to avoid forcing participants into a dichotomous choice.

The results suggest that misinformation susceptibility in this cohort is less about outright belief and more about uncertainty. Rather than confidently accepting falsehoods, many participants grapple with assessing veracity, demonstrating uncertainty in evaluating information rather than a firm belief in misinformation. While this highlights significant gaps in knowledge, it also suggests a form of intellectual humility, an awareness of their own limitations rather than unwarranted confidence in incorrect information. This openness to doubt creates opportunities for further learning.

Specifically, political and institutional claims, such as "The government is manipulating the public's perception of genetic engineering" (Item 1) and "The corporate media is controlled by the military-industrial complex: the major oil companies own the media and control their agenda" (Item 11), see low correct identification rates and high uncertainty. Similarly, health-related misinformation, such as "The government is knowingly spreading disease through the airwaves and food supply" (Item 2), sees low accuracy rates, pointing to potential vulnerabilities in public health misinformation discourse.

Rather than clear-cut misinformation belief, this cohort's susceptibility seems to be shaped by uncertainty and hesitation. The lack of confident rejection of false claims suggests that many young adults may be persuadable when faced with misleading narratives, making them more susceptible to misinformation influence over time. The accompanying graph illustrates how misinformation susceptibility manifests not just in belief, but in doubt.

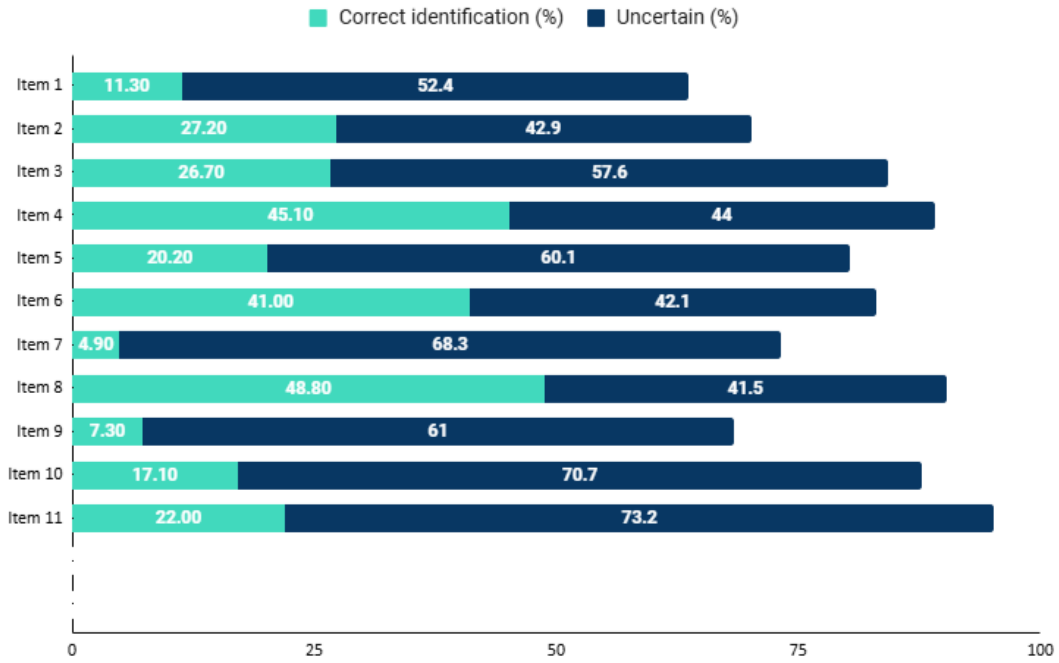


Fig. C6: Table showcasing participants' ability to discern misinformation headlines. This set of MIST items was not presented in the traditional 8-, 16-, or 21-item scale format and is therefore primarily exploratory. Some items were replaced to better fit the context, and as a result, the sample size varies across statements, ranging from 42 to 187 participants (average N = 114).

Young adults (N = 200) assessed 18 screenshot-based items from social media to gauge their ability to judge visual content reliability. While other measures allow us to maintain a controlled design and isolate specific factors (e.g., vocabulary used in headlines), this task requires participants to integrate multiple cues, such as source credibility, engagement metrics, and visual elements. As this is more closely aligned with how they encounter news in everyday life, it serves as a highly ecologically valid context for assessing veracity. Together, these different measures offer a more nuanced understanding of how young adults make sense of and judge the reliability of information they find online. The data reveals significant uncertainty, with uncertain responses often matching or exceeding correct identifications.

On average, only 41% of responses correctly classified the visuals, while uncertainty surpassed 50% on several items. This suggests that many young adults struggle to evaluate credibility, as assessing source reliability, engagement metrics, and visual cues requires critical analysis skills they may lack. The high uncertainty rate signals a lack of confidence in verification, where participants may recognise potential red flags but struggle to determine accuracy or seek reliable confirmation methods.

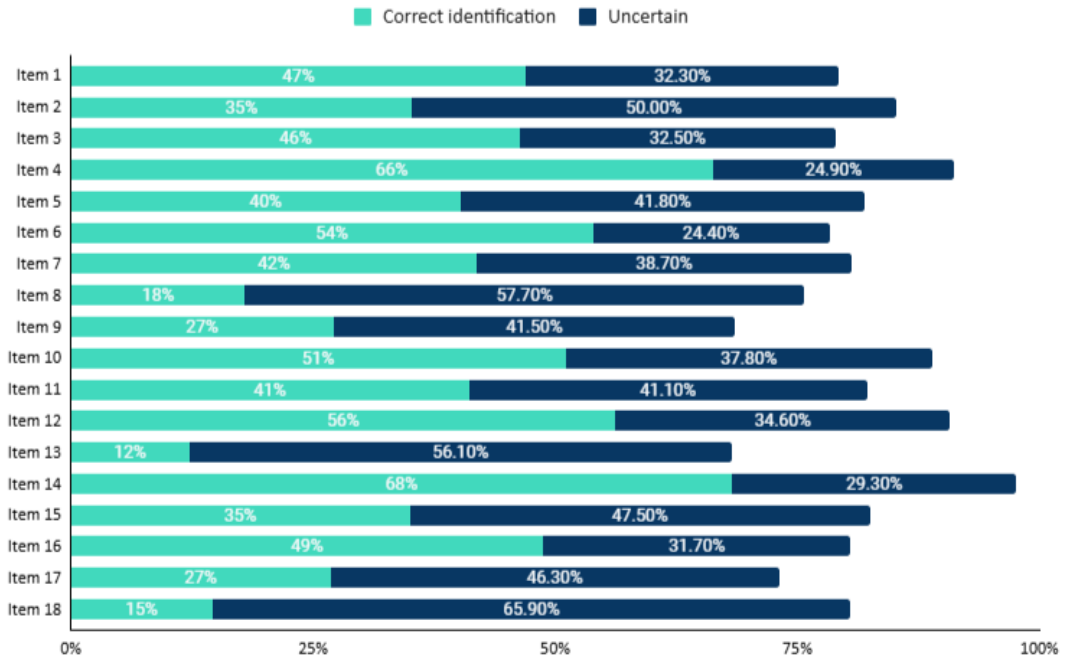


Fig. C7: Table showcasing participants' ability to discern veracity of visual content. Some items were replaced to better fit the context, and as a result, the sample size varies across statements, ranging from 41 to 200 participants (average N = 149).

3.2 Developing critical thinking

The following sections present insights into young adults' knowledge retention, reasoning abilities, and behavioural shifts after the workshops.

3.2.1 Significant improvements of key media literacy concepts

We assessed students' understanding of disinformation and confirmation bias at three time points: before, immediately after, and up to three months after the workshop. Our results suggest that after participating in Square's workshop, participants demonstrated significant improvements in their understanding of key aspects of legal rights and regulations and media literacy concepts.

Firstly, the results show a clear shift in how participants perceive social media's purpose. There was a notable increase in recognising its profit-driven nature (N = 233). Over time, perceptions partially reverted, though communication did not regain its initial dominance, and the view of social media as an informative tool increased (N = 66 - The follow-up sample was substantially smaller, reducing statistical power and potentially limiting the generalisability of long-term effects).

Secondly, our results suggest that after participating in Square's media literacy workshop, participants demonstrated significant improvements in their understanding of all key media literacy concepts (N = 164 to 398 depending on items - the improvement in disinformation definition was not statistically significant, p = 0.126). Notably, these gains persisted over time, with no significant decline observed at follow-up testing, but a considerable imbalance in sample sizes, due to attrition, may affect the robustness of the follow-up data (N = 67).

Our cohort also showed significant improvements in key aspects of legal rights and regulations. Several items showed increases in correct responses from pre-test to post-test (N = 234 to 407 depending on

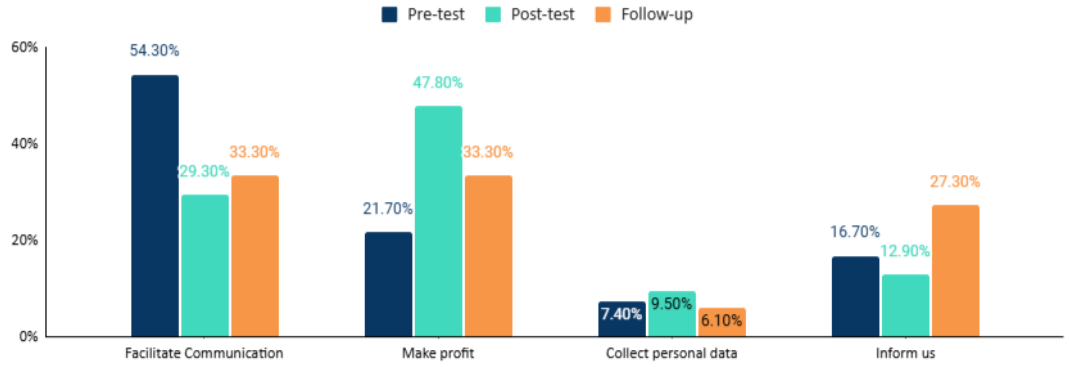


Fig. C8: Participants' perceived purpose of social media platforms before and after the workshops

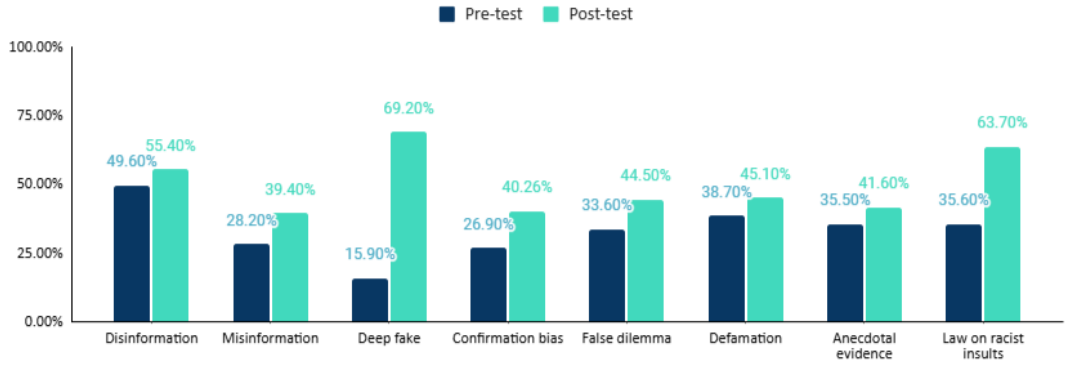


Fig. C9: Participants' ability to define key media literacy concepts before and after the workshops

items), paired with fewer "I don't know" answers, indicating that participants felt more confident in evaluating claims around law, free speech, and social media practices.

The follow-up data presents a more mixed picture (N = 69). Some statements, such as "The argument from authority aims to convince by using the origin of a statement rather than its content", maintained much of their post-test improvement, suggesting sustained learning. Others, particularly the one related to algorithmic filtering, saw a re-emergence of uncertainty at follow-up, in some cases exceeding pre-test levels. This could indicate that, despite initial gains, participants are finding these topics more complex than expected. The increase in uncertainty likely reflects the inherent difficulty of these issues, with participants feeling unsure or willing to admit to the complexity and lack of clarity around legal rights and platform regulations.

3.2.2 Enhanced fact-opinion differentiation over time

After participating in Square's workshop, participants showed notable progress, with significant improvements across most statements (N = 175 to 392 depending on items). Encouragingly, many of these improvements persisted over time, suggesting that the intervention had a lasting impact for at least 3 months (N = 70).

3.2.3 Gains in civic confidence amidst structural barriers

Square's workshop fostered a sustained increase in political confidence (N = 355), with gains not only maintained but, in some cases, further strengthened up to three months later (N = 73 at follow-up), empowering participants to engage more actively in political discussions and express their views with confidence. Participants reported feeling more knowledgeable about policy, more engaged in political discussions, and more capable of understanding and contributing to civic life after participating in Square's workshop. Notably, the strongest gains occurred in political self-expression, with more participants agreeing that they have political opinions worth listening to and something to say in debates. This suggests that the intervention helped foster a stronger sense of political voice and engagement and maintain it for at least three months after taking part in the workshop.

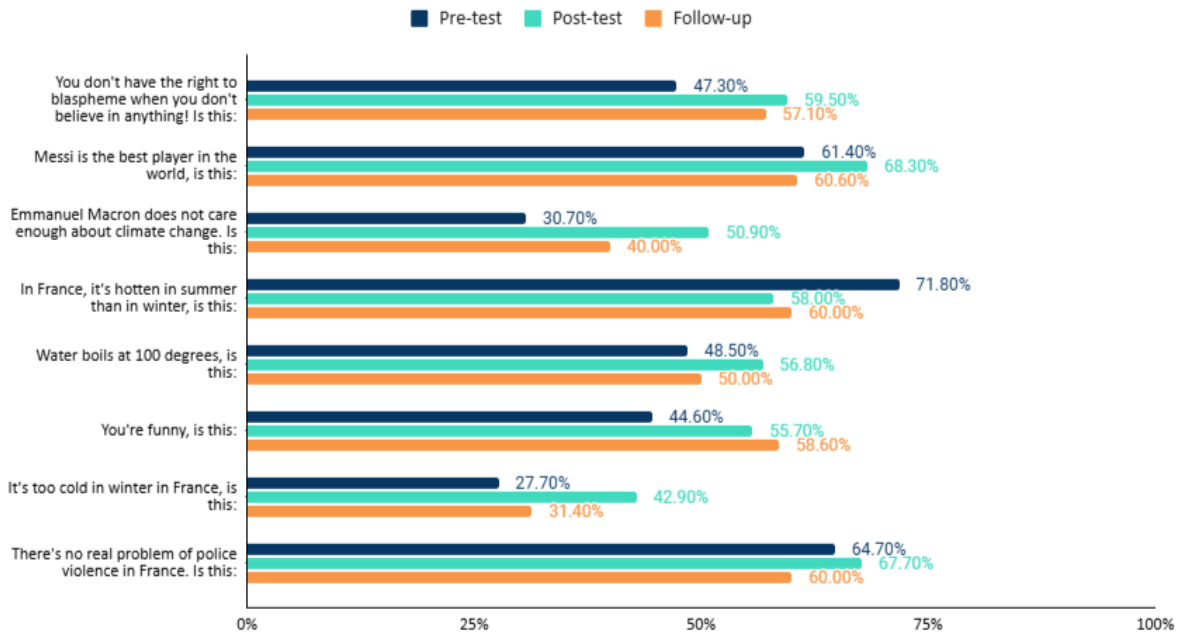


Fig. C10: Participants' ability to discern between facts and opinions, before and after the workshops

However, despite these improvements, a significant gap remains between the cohort and national averages. Even at follow-up, confidence levels fall well below national benchmarks, particularly in understanding political questions (31.5% vs. 54%) and feeling able to participate in political life (20.6% vs. 37%). These findings highlight that while progress has been made, deeper structural barriers to civic engagement persist, requiring long-term reinforcement and opportunities for practical political participation to close the gap.

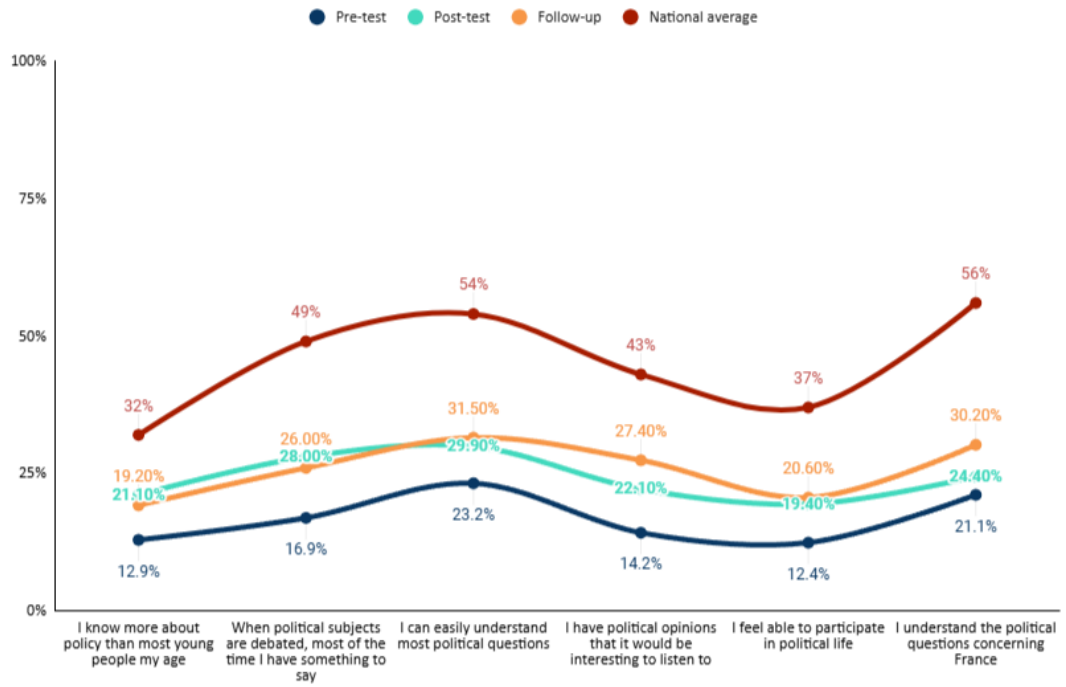


Fig. C11: Participants' civic confidence levels before and after the workshops, compared to the national average

3.2.4 Reduced conspiratorial thinking through targeted interventions

Average conspiratorial thinking scores showed a slight numerical decline, but our analyses found this change statistically insignificant. (N = 382)

In 2023, data showed a significant shift in attitude towards conspiratorial beliefs. Square's workshops lasted 20 hours over 4 days (against 2 days in 2024) and dedicated a specific session to conspiracy theories. The initial conspiracy level of this 2023 audience (recruited through local NGO channels) was much higher than the combined 2023/2024 ones (recruited by Mission Locales): 66.6% of the participants registered some level of agreement with ACBQ conspiratorial statements in 2023 against 40% in 2023/2024.

Out of the 228 participants in 2023, we focused on 90 individuals who completed all the key measures (i.e., conspiratorial thinking, Actively Open-minded Thinking, fact vs opinion discernment, theoretical disinformation knowledge) and at least 85% of the remaining scales in our questionnaires (N = 90). With an initial level of 66.6% of participants registering some level of agreement with conspiratorial statements, this figure declined by 16.6 percentage points post-workshop. This substantial reduction was both statistically significant ($p = 0.002$) and had a notable effect size (Cohen's $d = -0.6$), showcasing the effectiveness of the workshop in altering beliefs in conspiracy theories. This is aligned with a systematic review that assessed the effectiveness of various interventions in countering conspiracy beliefs and found that interventions fostering an analytical mindset or teaching critical thinking skills were notably successful in altering such beliefs (O'Mahony et al., 2023).

3.2.5 From intention to action: sustained behavioural changes

Participants were asked on their reflections on the workshop, offering insight into how they perceived its impact on their media literacy skills and online behaviours ($N = 183$ to 229 at post-test). The responses suggest that the training was effective in improving awareness and critical thinking, particularly in their perceived ability to distinguish fact from opinion and recognise manipulation on social media. Encouragingly, these self-reported gains align with the actual improvements observed in participants' ability to differentiate fact from opinion. Many also reported being more cautious about online content and more likely to check sources, indicating a shift towards more critical engagement with information.

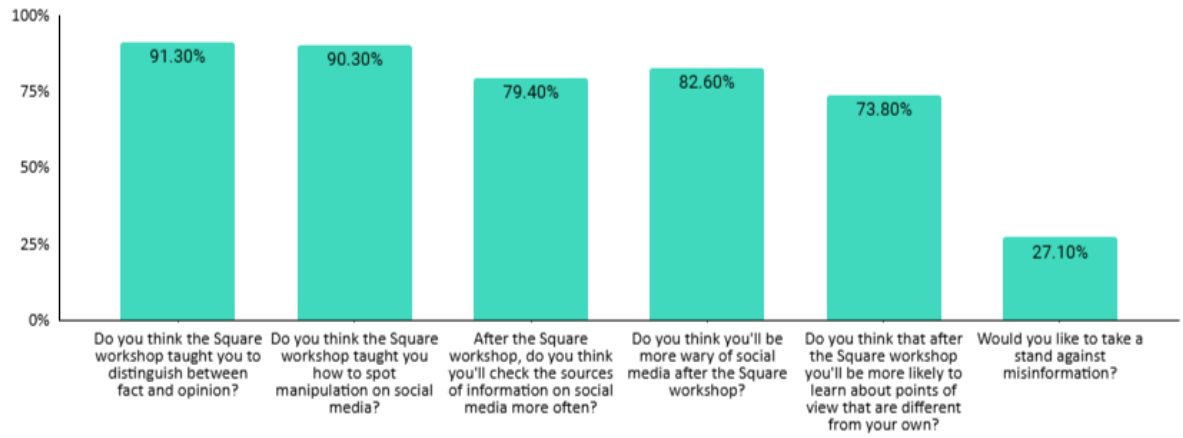

100%
75%
50%
25%
0%
91.30%
90.30%
79.40%
82.60%
73.80%
27.10%
Do you think the Square workshop taught you to distinguish between fact and opinion?
Do you think the Square workshop taught you how to spot manipulation on social media?
After the Square workshop, do you think you'll check the sources of information on social media more often?
Do you think you'll be more wary of social media after the Square workshop?
Do you think that after the Square workshop you'll be more likely to learn about points of view that are different from your own?
Would you like to take a stand against misinformation?

Fig. C12: Participants' post-workshop sentiments

Participants not only expressed a commitment to engaging more critically with online content after the workshop but also followed through on these intentions months later. The proportion of individuals who reported checking sources more frequently and seeking out diverse perspectives remained stable from post-workshop (N = 184) to follow-up (N = 61), indicating that these behaviours were successfully maintained over time. The most notable shift was a rise in scepticism toward news on social media, with more participants at follow-up expressing greater caution than they initially anticipated. These findings suggest that the workshop had a lasting impact on participants' digital habits, fostering critical engagement and greater awareness of misinformation risks. However, further research is needed to determine whether these self-reported behaviours lead to improved accuracy in assessing online content

4 Discussion

One of the greatest challenges in assessing impact was high attrition across testing points, which is typical for hard-to-reach populations and was particularly pronounced in this study. Distrust-driven survey engagement, participant drop-off, and difficulties in matching responses significantly reduced the number of complete cases available for longitudinal analysis. These disparities in sample sizes across time points, including between Q1 and Q2 for some variables, limited the ability to detect statistically significant effects, even when trends suggested meaningful shifts. Future work should focus on larger follow-up cohorts as well as incorporating control groups to ensure lower attrition and more robust findings.

The study also evolved over time, with adaptations to scales and metrics to better reflect local contexts and cognitive accessibility. While these adjustments improved relevance and accuracy, they complicated direct comparisons across testing waves. This means that instances of non-significance should be interpreted with caution, rather than as definitive evidence of no impact. Results tested during the 4-days intervention models exceeded those of the 1.5-day model, suggesting that longer workshops had greater effect. Despite these challenges, follow-up assessments revealed some retention of key media literacy concepts. Overall, follow-up surveys and longitudinal studies are critically needed in this space to enhance current understanding of media literacy and its long-term impact.

Despite these limitations, the report points to meaningful progress in several areas:

- Participants translated intention into action, maintaining their commitment to critical engagement with online content for up to three months.
- Participants reported feeling more engaged in political discussions and more confident in expressing their opinions post-workshop. The strongest gains were in political self-expression, with more participants believing they had opinions worth listening to. Future interventions should focus on long-term active engagement processes to foster positive results in engagement and political efficacy.
- Participants demonstrated clear improvements in defining key media literacy concepts, with gains in understanding confirmation bias, and algorithmic filtering.
- After the workshop, most participants reported checking sources more frequently, and sought out diverse perspectives, suggesting an increased awareness of verification practices.
- A shift in perception of social media's role was observed, with more participants recognising its commercial motives beyond being just a communication tool.
- Young adults showed improvements in their factual knowledge, with many gains sustained at follow-up.
- Confidence in judging the accuracy of information increased, with fewer "I don't know" responses on several misinformation-related items.

5 Conclusion

This report highlights the urgent need for targeted, evidence-based interventions in media literacy for disadvantaged youth in France.

Our findings reveal deep-rooted vulnerabilities in the ability of young adults to navigate today's complex information landscape. Across key measures, including institutional trust, civic confidence, media literacy concepts, and misinformation susceptibility, our participants demonstrate low accuracy, high uncertainty, and weak verification strategies, leaving them exposed to manipulation.

Additionally, institutional distrust further exacerbates conspiratorial thinking, a pattern clearly reflected in the data. This distrust correlates with a stronger belief in conspiracy narratives. As institutional scepticism deepens, so too does the reluctance to engage with credible sources, reinforcing a cycle where mistrust and conspiratorial beliefs fuel one another. The combination of institutional distrust, high levels of uncertainty and conspiratorial thinking suggests that many feel disconnected from civic life, question official narratives, and perceive themselves as excluded from meaningful discourse.

Motivated reasoning further complicates engagement with factual information. Despite direct exposure to media literacy training, a substantial proportion of participants continued to prioritise conviction over adaptability. This presents a fundamental challenge for media literacy interventions, as knowledge and skills can be taught, but the willingness to apply critical reasoning in the face of personal biases is much harder to instil. Square's workshops led to meaningful behavioural shifts, with participants

reporting increased source-checking and engagement with diverse perspectives. While these gains were more pronounced in some areas than others, the findings highlight both the potential of short-term interventions and the challenges of ensuring long-term retention.

This report reinforces that simply raising awareness about misinformation is not enough. Young people need cognitive tools, confidence, and motivation to actively engage in critical evaluation, source verification, and truth-seeking behaviours. The persistence of high uncertainty, reliance on surface-level verification, and low institutional trust signals deep-rooted vulnerabilities that cannot be addressed through short-term interventions alone. Instead of allowing the often-overlooked youth of France to fall deeper into distrust, conspiratorial thinking, and digital disengagement, we must empower them to think critically, assess information with confidence, and take an active role in society. Media literacy must be seen as an ongoing, structural need, requiring long-term investment, reinforcement, and adaptation to the evolving digital landscape. If left unaddressed, these gaps will only widen, exacerbating social divisions, political disengagement, and misinformation vulnerabilities. However, with targeted interventions, the marginalised youth of France can be equipped not just to navigate the digital world but to help shape it for the better.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

CEVIPOF (2021) Baromètre de la confiance politique, vague 12. https://www.sciencespo.fr/cevipof/fr/etudes-enquetes/barometre-confiance-politique

CEVIPOF (2024) Baromètre de la confiance politique. https://www.sciencespo.fr/cevipof/fr/etudes-enquetes/barometre-confiance-politique

CNESCO (2018) Enquête 'ecole et citoyenneté'

De Keersmaecker J, Roets A (2017) 'fake news': Incorrect, but hard to correct. the role of cognitive ability on the impact of false information on social impressions. Intelligence 65:107–110. 10.1016/j.intell.2017.10.005

Edwards L, Obia V, Goodman E, et al (2025) Cross-sectoral challenges to media literacy

Guay B, Berinsky A, Pennycook Gea (2023) How to think about whether misinformation interventions work. Nat Hum Behav 7 pp 1231,1233. 10.1038/s41562-023-01667-w

Haran U, Ritov I, Mellers BA (2013) The role of actively open-minded thinking in information acquisition, accuracy, and calibration. Judgment and Decision Making 8(3):188–201. 10.1017/S1930297500059217

Jolley D, Douglas KM, Skipper Y, et al (2021) Adolescent conspiracy beliefs questionnaire (acbq). APA PsycTests; 10.1037/t82252-000

Kovic M, Füchslin T (2018) Probability and conspiratorial thinking. Applied Cognitive Psychology 32:390–400. URL https://api.semanticscholar.org/CorpusID:148906494

Krosnick JA, Holbrook AL, Berent MK, et al (2002) The impact of 'no opinion' response options on data quality: Non-attitude reduction or an invitation to satisfice? Public Opinion Quarterly; 66(3):371–403. 10.1086/341394

Nera K (2023) Complotisme et quête identitaire. PUF, Paris

O'Mahony C, Brassil M, Murphy G, et al (2023) The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. 18(4). 10.1371/journal.pone.0280902

Pennycook G, Cheyne J, Koehler D, et al (2020) On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. Judgment and Decision Making; 15(4):476–498. 10.1017/S1930297500007439

Pfänder J, Altay S (2025) Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. Nat Hum Behav 9 pp 688,699. 10.1038/s41562-024-02086-1

Potter J (2022) Analysis of definitions of media literacy. Journal of Media Literacy Education 14(2):27–43. 10.23860/JMLE-2022-14-2-3

Roozenbeek J, van der Linden S, Goldberg B, et al (2022) Psychological inoculation improves resilience against misinformation on social media. Science Advances 8(34):eabo6254. 10.1126/sciadv.abo6254

Schulz W, Ainley J, Fraillon J, et al (2010) International civic and citizenship education study international survey

Stanovich K, Toplak M (2023) Actively open-minded thinking and its measurement. J Intell pp 11,27. 10.3390/jintelligence11020027

AI, Digital Democracy, and the Crisis of Social Cohesion: Rebuilding Political Trust

Adrien Tallent[1]

[1] Sorbonne University, Paris, France.

Corresponding author. E-mail: adrien.tallent@gmail.com

Abstract

In this paper, I examine the ambivalent role artificial intelligence (AI) and digital technologies play in contemporary democracies. While AI-driven systems can intensify polarization, obscure accountability, and erode public trust, they also hold potential to strengthen democratic participation, social cohesion, and deliberation. I argue that the impact of AI is not predetermined: it depends on how technologies are embedded in institutional frameworks and political cultures.

I distinguish between two approaches: (i) digital technology as a tool for democracy that can enhance public debate, enable pluralism, and support civic engagement through platforms like vTaiwan or Decidim; (ii) and digital technology as a substitute for democracy that automates weak consensus and decision-making. Drawing from Rousseau's concept of general will and Habermas' theory of communicative action, I explore the risks of delegating political reasoning to machines and the dangers of algorithmic consensus that bypass human conflict and deliberation.

Therefore, I discuss alternatives such as pluralism-aware recommendation algorithms, emotionally intelligent moderation tools, and context-sensitive interventions to restore deliberative norms. Finally, I argue that rebuilding democratic trust in the age of AI requires rethinking institutional architectures. This involves not only decentralizing power, but embedding civic participation and democratic oversight within governance structures. Only by integrating these practices into accountable institutions can we effectively counter the dominance of Big Tech and ensure that technological development serves democratic ends rather than undermining them.

Keywords

Digital democracy, Habermas machine, Rousseau, Institutions, Trust.

1 Introduction

Artificial intelligence (AI) and data-driven systems have become increasingly central to the organization of public and political life (O'Neil, 2016), extending well beyond traditional communication technologies and content curation, insofar as they reshape the modalities of governance, the structure of democratic participation, and the epistemic foundations of collective deliberation. However, the role of AI in democratic societies is ambivalent. On the one hand, algorithmically driven platforms contribute to political polarization, the spread of misinformation, and a generalized erosion of civic and institutional trust (Sunstein, 2018; Tucker, Guess, Barbera, Vaccari, Siegel, Sanovich, Stukal, and Nyhan, 2018). On the other hand, research shows that AI has the potential to promote more inclusive forms of public engagement, improve deliberative processes, and increase public engagement (Duberry, 2022; Helbing, Frey, Gigerenzer, Hafen, Hagner, Hofstetter, van den Hoven, Zicari, and Zwitter, 2019; Landemore, 2020).

"Political trust" refers to citizens' confidence in formal political institutions (e.g., parliaments, courts, ministries) and in the procedures that authorize them. Political trust is "the glue that keeps the system together"(Van der Meer and Zmerli, 2017). Political trust is therefore crucial to representative democracies. By contrast, "public trust" refers to confidence in the public sphere (Habermas, 1984) as an arena of communication (trust in media infrastructures, platforms, and in fellow citizens as interlocutors). In short, political trust concerns authority and accountability, while public trust concerns the credibility, inclusiveness, and fairness of discourse conditions.

At the heart of these debates is a fundamental tension between optimization and democratic legitimacy. Contemporary algorithmic systems are built on principles of efficiency, prediction, and behavioral nudging – what Morozov (Morozov, 2013) describes as "technological solutionism." These systems often prioritize calculability and speed over transparency, contestability, and participation. As governance becomes increasingly mediated by data infrastructures, key political decisions – including welfare eligibility (Rachovitsa and Johann, 2022), predictive policing (Meijer and Wessels, 2019), and urban planning (Sanchez, Brenman, and Ye, 2025) – are delegated to automated systems that frequently operate as "black boxes" (Pasquale, 2016). In such cases, political authority shifts from deliberative institutions to technical expertise and opaque algorithms, raising critical concerns about accountability, democratic oversight, and normative legitimacy (Eubanks, 2018; Yeung, 2018). Today, public opinion is largely formed in digital spaces, where algorithmically-curated content governs what citizens see, how they engage and interact (Perra and Rocha, 2019; Shearer, 2021).

In this paper, I examine the dual role played by AI and digital technologies in contemporary democracies, particularly on the dynamics of trust, social cohesion, legitimacy, and public trust. While AI is often proposed as a neutral instrument of efficiency, it is increasingly clear that its deployment is neither politically nor normatively neutral. Recommendation systems and predictive algorithms have reshaped the epistemic foundations of democratic life by prioritizing visibility and engagement, over accuracy, argumentation, deliberation, and trust (Dijck, Es, Helmond, and Vlist, 2025). These developments unfold within a broader context of democratic fragility. Representative democracy faces a crisis of legitimacy fueled by long-standing socioeconomic inequalities, political disengagement, and rising populist sentiment (Rosanvallon, 2013; Tormey, 2014). Although these phenomena are not reducible nor solely due to technological change, digital infrastructures, including AI systems, have contributed to their intensification. Automated decision-making, for example, introduces new forms of opacity, personalized content feeds reinforce cognitive and ideological closure (Deffains, 2022; Yeung, 2018), and the delegation of agenda-setting, visibility and moderations rules for public conversation to private platforms by Big Tech companies (i.e. large technological firms that control digital infrastructures and services which structure access to the public sphere) weakens trust in democratic institutions (Zuboff, 2019).This creates a de facto privatization of this public conversation. Within this context, the notion of "digital democracy" (Congge, Guillamón, Nurmandi, Salahudin, and Sihidi, 2023) emerges as both a normative aspiration and an empirical ambiguity. For Novelli and Sandri (Novelli and Sandri, 2024), this notion "emerges from blending democratic governance with computer-mediated communications" (p. 1) and therefore implies a major political change that could mean enhancing or at least transforming democracy.

The central thesis of this paper is that despite the contribution to the weakening of political trust, AI nevertheless holds the potential to support more participatory and deliberative forms of democratic engagement. The paper is structured as follows. First, I contextualize the political risks of AI in relation to the historical project of rational governance. Second, I explore how AI can be redirected to support democratic participation and pluralism. Third, I propose institutional reforms for embedding AI within democratic infrastructures.

2 Context: AI against democracy

The political framework of Western democratic societies is largely supported by the concept of the social contract (Rosanvallon, 2013). While key theories from Rousseau (Rousseau, 2018) and Hobbes (Hobbes, 1985) differ, they share a common goal: understanding why rational individuals would choose to give up some personal freedoms in exchange for social and political order.

Within the social contract, and more broadly within Enlightenment philosophy, two concepts seem to be key to understanding today's political landscape. First, the idea of progress—especially as articulated by Condorcet (Condorcet, 1822)—links technological advancement to human development. Second, the emphasis on reason as the unique human faculty leads to searching for the most rational way to govern individuals and the political society emerging from the contract. Hobbes' Leviathan (Hobbes, 1985), for example, describes society as a machine governed by mathematical rules, reflecting a mechanistic and rationalist view of political order that continues to influence contemporary thought. This rationalist legacy is later echoed in cybernetics, which applies principles of control, feedback, and information processing to both machines and societies, reinforcing the idea that governance can be automated (Wiener, 2019). AI as a mode of governance can thus be seen as part of a history of the mathematization of political power. However, this "governance by numbers" (Supiot, 2015) is fundamentally a political choice aimed at reducing the act of governing by transforming it into "algorithmic governmentality" (Rouvroy and Berns, 2013). Here, decisions are based on data processing and calculation rather than on political deliberation or social norms considered inefficient, slow, or hazardous. In doing so, AI is both the embodiment and the victory of the cybernetic project that precedes and supports it (Hunyadi, 2025).

By reducing reality to exploitable parameters and giving algorithms a central role in managing behavior at a "micro-level" (Turlakova and Lohvinenko, 2023), AI does not merely assist public action, it profoundly redefines it. The AI we know today is a statistical, predictive object, part of a process of reducing reality to a few parameters, which masks the political and ideological choices – human choices – embedded in its design. This vision of technological neutrality then serves a political project of a managerial and algorithmic vision of public decision-making (Mowshowitz, 1984). The rise of AI is thus part of the long history of the political and technical dream of a perfect political order (Rouvroy and Berns, 2013; Supiot, 2015). This long-standing dream aims to bypass the individual in order to achieve a regulated and optimized society, governed not by law, but by data, algorithms, and predictive models. Furthermore, the use of AI for authoritarian purposes, e.g., social credit scoring in China (Creemers, 2018), or at least anti-democratic purposes, e.g., DOGE in the US (Kelly, 2025), reflects a political shift made possible by technology as a way of reducing in the scope of the social contract.

However, nothing is inevitable. While technological advances like AI pose significant challenges, it would be reductive to treat them as the root causes of democratic decline. These challenges are not technological *per se*, but political. As Feenberg (Feenberg, 2012) argues, technologies are not neutral tools but reflect the values and power relations of the societies producing them. The rhetoric of inevitable innovation frames AI as beyond democratic control, obscuring the fact that algorithmic systems are designed and governed by humans who make normative choices. Issues such as algorithmic bias, opacity, and exclusion (Eubanks, 2018; Noble, 2018; O'Neil, 2016) illustrate that data and AI are inherently political. The responsibility, therefore, lies not with the machines themselves, but with the institutions and actors who design, deploy, and regulate them.

3 How digital technologies can contribute to rebuild social cohesion and facilitate deliberation

In the following sections, I will address various technological initiatives that may be used to strengthen democracy. When mentioning "AI", I distinguish two families: (i) ML-based recommender/decision systems (e.g., ranking and moderation); and (ii) large language models (LLMs) used as mediators or assistants in deliberation. When I discuss platforms, voting portals, and participation tooling in general, I refer to digital technologies.

On the positive, digital technologies have enabled new democratic experiments, including civic platforms, participatory budgeting tools, and AI-assisted deliberative forums (Farrell and Suiter, 2021; Landemore, 2020). These initiatives suggest that under specific normative and institutional conditions, AI can support democratic renewal rather than its erosion. This potential, however, is neither automatic nor guaranteed and depends on how AI is designed, governed, and embedded within broader political arrangements. Burgess (Burgess, 2022) suggests four ways digital technologies can be used to enhance

democracy: (1) the transfer of voting for representatives online; (2) online voting to pass / reject bills; (3) the use of individuals preferences to inform legislative decision making; (4) the replacement of the (physical) legislature and individual representatives . With this understanding, digital technologies can impact all types of democracy: direct, representative (elections), and deliberative (consultations, participatory decision-making).

However, I distinguish two approaches to better understand this impact: (i) digital technologies *for democracy*, which strengthens the capacity for collective action and values democratic conflict as a resource; and (ii) digital technologies *as a substitute to democracy*, which aims to automate choices by reducing the uncertainty of debate by smoothing it over and making it manageable. Key developments are initiatives such as vTaiwan (Hsiao, Lin, Tang, Narayanan, and Sarahe, 2018), Decidim (Aragón, Kaltenbrunner, Calleja-López, Pereira, Monterde, Barandiaran, and Gómez, 2017), and more recently the "Habermas Machine" study by Google DeepMind researchers designed to mediate online deliberations (Tessler, Bakker, Jarrett, Sheahan, Chadwick, Koster, Evans, Campbell-Gillingham, Collins, Parkes, Botvinick, and Summerfield, 2024). However, the latter raises concerns as it seeks to automate democracy rather than cultivate genuine political engagement. This reflects a broader tension in deliberative democracy between approaches that aim to construct consensus and those that enable contestation and structured disagreement.

3.1 Digital technologies *for democracy*

Internet and social media platforms were once seen as the ideal digital tools for promoting democracy (Abiteboul and Cattan, 2022) (Tréguer, 2023). Yet over time, they have often fallen short of these utopian expectations, giving rise instead to new forms of polarization, surveillance, and disinformation (Durand, 2020). Today, the rise of AI presents a renewed opportunity to reflect on the democratic potential of digital technologies. In recent years, political initiatives have been flourishing that use digital tools to revitalize representative democracy. Following in the footsteps of "civic hackers" and Audrey Tang (Frenkiel, 2017), who served as minister without portfolio in 2016 and then as the first Minister of Digital Affairs of Taïwan from 2022 to 2024. After the 2014 "Sunflower Movement" when young people challenged the government, Taiwan transformed its democratic system to include many e-participation software such as Pol.is, a "digital platform for opinion collection, to facilitate large-scale conversations and consensus building" . This led to the creation of vTaiwan, or "virtual Taiwan," a platform for citizen deliberation and co-creation of laws between citizens and legislators (Hsiao et al., 2018; Schneider, 2019). Similarly, the victorious *indignados* at Barcelona City Hall – organized as *Barcelona en Comú*, a political party born out of the 2011 *Indignados* movement – behind Ada Colau launched the Decidim.Barcelona platform in 2016 – based on the online participation software Decidim – to co-construct the municipal action plan for 2016-2019 (Aragón, Kaltenbrunner, Calleja-López, Pereira, Monterde, Barandiaran, and Gómez, 2017; Peña-López, 2019). While these are just a few of the many example of "civic tech" that offers participatory digital tools in the political sphere, we see here the positive argument that digital technologies can be mobilized to structure debate, rather than replace it, and encourage both agreement and disagreement in the populace, consequently promoting collective reflexivity.

Those platforms not only illustrate real-world success in using technology for democratic participation, but also work to rebuild political trust by making institutional processes more accessible, deliberative, and transparent. Political trust, i.e. the confidence citizens place in institutions, decision-makers, and democratic procedures, is a foundational element of democratic life (Easton, 1967). Trust is both sensitive to perceptions of competence and fairness, and becomes fragile when institutions are opaque or unaccountable (Bouckaert and van de Walle, 2003). Research suggests that trust can be repaired when citizens feel heard, see their contributions reflected in outcomes, and witness fair procedural conduct (Ananny and Crawford, 2018; Grimmelikhuijsen, Porumbescu, Hong, and Im, 2013). Digital tools, if designed with these values in mind, can enhance not just participation, but the credibility and perceived legitimacy of democratic institutions themselves. I will discuss this in the final section.

3.2 AI *as a substitute to democracy*: the case of the "Habermas machine"

In this sub-section, I examine the use of large language models (LLMs) to assist deliberation. Inspired by Jürgen Habermas' theory of communicative action (Habermas, 1984,8), the "Habermas Machine" (Tessler et al., 2024) aims to facilitate democratic deliberation by helping groups reach consensus on complex and divisive issues. It functions as a virtual mediator, using advanced language models to analyze individual opinions, generate statements reflecting diverse viewpoints, and produce consensus statements

based on participant feedback. The goal is to maximize collective approval while fairly incorporating all perspectives expressed. According to their results, participants seem to prefer AI-generated statements to those written by human mediators, judging them to be "more informative, clearer, and more impartial." After the deliberation period, participants often updated their views and thus converged toward a common perspective.

While it has *prima facie* laudable intentions, this project can be considered as anti-democratic because it seeks to reduce conflict, smooth out opinions, and automate the production of consensus. DeepMind researchers praise the speed and efficiency of their system compared to human deliberations, which they consider slow and inefficient (Tessler et al., 2024). Importantly, Habermas himself distanced himself from the project (Rabe, 25. April 2025). For him, the process of conflict resolution cannot be delegated to technology since deliberation is based on each participant being able to sensitively adopt the other's point of view. The experiment reflects a technocratic and managerial view of democracy, that is to say reducing democracy and political deliberation to a process of optimization and problem solving. There is a risk of the emergence of a form of citizenless democracy – democratic procedures continue to operate but without meaningful citizen participation. Why bother discussing and arguing when AI can instantly produce a consensus? This approach may deny the fact that the democratic process is not only about making decisions, but also about forging a political community through debate and collective action. By systematically seeking consensus and reducing the complexity of human situations to algorithmic variables, it risks erasing the nuances and uniqueness of each individual and reducing the scope of the political contract.

This invites a deeper reflection: should AI systems aim to secure consensus, or should they leave space for disagreement and political opposition as essential features of democratic life? To explore this tension, it is necessary to turn to the philosophical foundations that inform different conceptions of collective decision-making.

3.3 Theoretical Tension: Consensus or political opposition?

I consider two different theoretical foundations for designing an AI model to support democracy: Habermas' ideal of rational consensus and Rousseau's concept of the general will present distinct visions of collective decision-making, with important implications for how such AI system should be designed.

3.3.1 Habermasian Rational Consensus

Habermas's theory of communicative action offers a procedural model of democratic legitimacy grounded in rational discourse. At its core is the concept of the ideal speech situation, a hypothetical communicative setting in which participants are free from coercion and engage as equals, each capable of raising and responding to validity claims through reasoned argumentation (Habermas, 1984,8). Under these conditions, consensus is not imposed but emerges organically through mutual understanding. However, Habermas frames consensus not as an expected outcome, but as a regulative ideal that guides deliberative practices and normative critique (Baynes, 1991; Bohman and Rehg, 2014). This emphasis on deliberation supports a vision of democracy rooted in communicative rationality.

3.3.2 Rousseau's General Will

Rousseau's concept of the general will designates a unified expression of the collective interest transcending the aggregation of individual preferences. Rousseau emphasizes that the general will is not reducible to the will of the majority; rather, it is what "tends always to the preservation and welfare of the whole" (Rousseau, 2018). The legitimacy of law, in his view, stems from its alignment with this collective interest, and not from mere aggregated agreement. Accordingly, genuine participation in the general will demands that citizens abstract from their particular interests and deliberate from the standpoint of the common good (Bertram, 2004; Cohen, 2010; Tallent, 2025). This normative ideal introduces a strong conception of civic unity that resists both factionalism and pure interest bargaining. Yet, Rousseau also acknowledged the necessary role of dissent and pluralism in shaping and refining the general will, especially through open political deliberation where conflicting views are confronted and particular interests are transcended into a higher expression of the common good (Marks, 2005).

3.4 Design Implications

Designing AI for democracy first requires asking what kind of democratic ideal such systems should serve. This question motivates the discussion of two contrasting normative models: one centered on

consensus, the other on pluralism. Modern AI systems face a tension between these two accounts, and therefore a key design challenge lies in navigating between them. This section does not aim to resolve the tension between competing normative models, but rather to clarify the stakes by outlining two contrasting visions that inform how democratic legitimacy might be encoded into AI systems.

Given these two key theories that aim to determine political legitimacy in democracies, we can derive two models that AI systems could adopt to support democracy:

i) A Consensus-Oriented AI, broadly aligned with Habermasian ideals of communicative rationality, aspires to enhance deliberative quality by fostering inclusive, reasoned dialogue. Such systems may include real-time suggestion tools that encourage civility, evidence-based contributions, and the clarification of argumentative structure through logic-based mapping or discourse modeling (Binns, Van Kleek, Veale, Lyngs, Zhao, and Shadbolt, 2018; Hovy and Prabhumoye, 2021). These interventions could support mutual understanding and reciprocal engagement. However, critics warn that emphasizing procedural harmony risks marginalizing legitimate dissent or prematurely closing off contestation, thereby instituting a form of artificial consensus that undermines political plurality (Morley, Floridi, Kinsey, and Elhalal, 2020). The danger of this type of consensus is its weakness and the possible reactions of human-agents trying to "hijack" the process by manipulating inputs in order to make the mediator lean towards their position. It also risks imposing artificial harmony that suppresses legitimate conflict.

ii) By contrast, a Pluralism-Preserving AI, drawing inspiration from Rousseauian concerns about political homogeneity and forced agreement, could maintain spaces of disagreement and political contestation. Rather than resolving differences, such systems might adopt distributional approaches, e.g., exposing users to a balanced range of perspectives or tailoring outputs to reflect population-level value diversity (Draws, Inel, Tintarev, Baden, and Timmermans, 2022; Helberger, Karppinen, and D'Acunto, 2018). These models resist privileging any single interpretation of the "common good" and instead seek to preserve normative heterogeneity in pluralist societies. However, without mechanisms to mediate between conflicting views, such an approach risks developing polarization and may, paradoxically, expose minorities to the soft tyranny of majority preferences if not carefully safeguarded.

These two models show how AI might support democracy, but they leave open an important question: how can people stay together when consensus is weak and pluralism creates division? The next section looks at other models that try to answer this question beyond the idea of public debate.

4 Alternative models for social cohesion

Social cohesion refers to the strength of relationships and a sense of solidarity among members of a society, encompassing how people interact, trust each other, and share common values and goals. Following Rousseau, private actors are unable to embody the general will as they necessarily represent private interests and thus oppose public interest (Rousseau, 2018).

In this section, I explore three possible directions for alternative algorithmic models. Alternative recommendation algorithms can prioritize diverse perspectives to help mitigate the polarizing effects of traditional engagement-based ranking models. They can also similarly provide real-time AI interventions for de-escalating online hostility to promote more respectful and empathetic interactions. Lastly, context-aware intervention mechanisms could reduce the prevalence of online hostility and fostering deliberative norms.

4.1 Alternative recommendation algorithms that prioritize diverse perspectives

AI recommendation systems (learning models used to rank, filter, and personalize exposure to content) often reinforce epistemic bubbles by maximizing engagement metrics, e.g., clicks, views, and time spent, leading to personalization loops that prioritize popular and/or emotionally-charged content (Schrage, 2020). This feedback loop has shown to foster polarization, conspiracy thinking, and hostility in online environments (Bozdag, 2013). Helberger et al. (Helberger et al., 2018) argue for the need of "pluralism-aware" recommender systems that proactively expose users to diverse, high-quality content, aligned with democratic values like reason-giving, mutual respect, and informed discourse.

One recent study provides strong empirical grounding for the potential of personalized AI interactions to counter misinformation, reduce harmful beliefs in conspiracies, and encourage reflective thinking (Costello, Pennycook, and Rand, 2024). Using GPT-4 Turbo, the study placed 2,190 participants

identified as conspiracy theorists in three-round conversations with the AI model with evidence-based, belief-tailored counterarguments. The results showed approximately 20 percent decrease in belief strength that persisted for at least two months. Participants also exhibited reduced belief in unrelated conspiracy theories and reported feeling more informed and less anxious.

Several algorithmic innovations have also been recently introduced to address the popularity bias and, therefore, to improve the diversity of recommended content. *FairMatch* (Mansoury, Abdollahpouri, Pechenizkiy, Mobasher, and Burke, 2020), for example, increases aggregate diversity by recommending underrepresented yet high-quality items. DGCN (Zheng, Gao, Chen, Jin, and Li, 2021) integrates diversity via graph convolutional networks, while hybrid systems (Xie, Du, Zhao, Liu, and Li, 2021) optimize for multiple objectives such as novelty and coverage. The "Leave No One Behind" framework (Li and Zhou, 2025) further advances socially aware recommendation. A 2023 survey (Zhao, Wang, Liu, Cheng, Aggarwal, and Derr, 2024) synthesizes these advancements and emphasizes the importance of embedding values such as equity, representation, and exposure to diversity, particularly in domains that shape public opinion and civic life.

Applied to political trust and participation, careful experimental metrics of deliberative quality (such as mutual understanding and justification) could complement accuracy-based evaluation in order to support public reasoning and democratic renewal. However, such measures risk reducing discursive practices to simplistic indicators that are sensitive to cultural context and easy to "optimize" without any real improvement. They can also induce performative effects (talking to tick boxes), shift attention from ends to measurable means, and increase the need for monitoring and evaluation to the detriment of deliberation itself. Therefore, such indicators should remain advisory signals with human review procedures, and not optimization targets.

4.2 Real-time AI interventions for de-escalating online hostility

A second direction could be the careful use of AI interventions for de-escalating online hostility to promote more respectful and empathetic interactions in digital spaces. For instance, Gillespie (Gillespie, 2018) and Jhaver et al. (Jhaver, Birman, Gilbert, and Bruckman, 2019) emphasize the role of content moderation in shaping online discourse, and highlight how emotionally attuned interventions, whether automated or human-led, can help mitigate harmful interactions.

Empirical studies confirm the effectiveness of such strategies. Chang et al. (Chang, Schluger, and Danescu-Niculescu-Mizil, 2022) demonstrate that proactive, context-aware interventions, such as prompting users to reflect, slow down heated exchanges, or by emphasizing shared values, can significantly reduce conflict. Similarly, Ye et al. (Ye, Wang, and Xu, 2025) show that contextual AI interventions targeting hostile attribution bias improve emotion regulation and reduce aggression, particularly among adolescents. Emotion-regulation tools and conflict-resolution systems, such as Happy Software (Filella and Ros-Morente, 2023), have shown that structured prompts can increase empathy and assertiveness in digital communication. Importantly, the normative aim of these interventions extends beyond basic civility. As Dryzek (Dryzek, 2002) argues in his theory of deliberative democracy, the goal is not only to avoid hostility but to foster a culture of deliberative ethics—encouraging mutual understanding, reason-giving, and respectful disagreement.

However, the attempt to use AI systems to generate constructive and more friendly dialogues and opinions is not immune to the criticism I have formulated against the "Habermas Machine," namely a reduction in the diversity of viewpoints by favoring conciliatory or acceptable formulations for the majority. While it may be true that AI mediation can ease conflicts and increase reported satisfaction levels, these gains may also reduce the range of claims on controversial issues. Consequently, any development of this type of solution should combine strong theoretical and practical requirements based on principles of pluralism rather than aiming solely at optimizing agreements (Argyle, Busby, Gubler, Bail, Howe, Rytting, and Wingate, 2023).

4.3 Context-aware intervention mechanisms to reduce the prevalence of online hostility and promote deliberative norms

Context awareness is a concept used to describe AI systems capable of adapting to each user—their environment and needs–in order to provide personalized recommendations (Hong, Suh, and Kim, 2009). More specifically, Hong et al. (Hong et al., 2009) define user context as "any information that can be used to characterize the situation of an entity" (p. 1).

So what does this mean in the context of democracy and political trust? Rather than applying uniform moderation strategies, community-sensitive approaches can tailor interventions to the specific social norms and behavioral expectations of distinct online spaces. A large-scale study found that such community-adaptive moderation significantly improves outcomes in reducing incivility and promoting more cooperative dialogue (Park, Mendelsohn, Radhakrishnan, Jain, Kanakagiri, Jurgens, and Tsvetkov, 2021). For instance, 'soft prompts' encouraging empathy, such as reminders to consider others' perspectives or emotions, have been shown to increase pro-social behavior (Park, Yim, Chung, and Lee, 2023). Ye et al. (Ye et al., 2025) concur and argue that positive social interventions can indeed reduce hostility and promote a climate of empathy between adolescents online

However, this kind of experiment raises the question of the standard against which messages and expressions are compared. Systems that classify "toxicity" or "civility" too often incorrectly flag non-standard dialects spoken by minorities: for example, they produce more false positives for African American English (Sap, Card, Gabriel, Choi, and Smith, 2019). As a result, already marginalized speakers risk being over-penalized. Any use of these signals must therefore include bias audits by dialect/identity, metrics independent of the decision threshold, and human appeal processes. Moreover, the normative objective of these practices should go beyond enforcing civility as a minimal behavioral standard. Following Dryzek (Dryzek, 2002), democratic discourse must be underpinned by deliberative ethics, i.e. norms that promote mutual justification, reciprocal listening, and critical reflection. Context-aware moderation tools can help instantiate these ideals within digital infrastructures, gradually shifting expectations and behaviors toward more constructive public engagement. By embedding deliberative values into platform design, these interventions could support a more participatory, equitable, and reflective democratic culture. Such mechanisms can thus be seen as part of a broader democratic infrastructure project and a way re-legitimize public discourse by offering an opportunity to rebuild political trust from the bottom up: not through surveillance or coercion, but encouraging participation and shared accountability.

5 Rethinking institutional infrastructures

Finally, to ensure that AI contributes to democratic trust and social cohesion, it is crucial to rethink institutional governance structures, particularly in response to growing concerns over Big Tech's concentrated power, which undermines both. Many critiques focus on decentralization as a solution to Big Tech dominance (Durand, 2020; Fenwick and Vermeulen, 2019; Kaal, 2021). In this final section, I argue that a key issue for combatting Big Tech dominance is the institutionalization of those initiatives by democratic institutions. The role of institutions in AI governance is central: instead of aiming to simply redistribute technological power from Big Tech to smaller actors, the challenge lies in redesigning the institutional infrastructures of our democracies. The challenge of this section is therefore to offer ways to "institutionalize" the democratic practices we have discussed in this article.

Institutions have been using AI technologies for some time now: since the 2000s, amid ongoing debates and questions about the implications of these technological transformations for public policy, many governments have been taking advantage of new information technologies (West, 2005). This evolution has not only influenced how institutions operate, but has also begun to alter the institutional frameworks through which democratic participation and authority are structured. Yet, the issue is not only the decentralization of power that is often highlighted – including by the tech barons themselves (Schmidt and Cohen, 2013) – when we talk about technology, but also the institutional transformation of democracy to integrate citizens' uses of technology.

Important to note, an institution exercises power over a territory. Traditionally, Western democracies based on the Social Contract are partly (some more than others) aligned with the logic of the nation state (Ménissier, 2020). In the case of AI and digital technology, a new space has emerged in this context: cyberspace. This particular type of space challenges traditional representations of the state (it is often said that digital technology "deterritorializes" politics (Lassègue and Garapon, 2021). The cyberspace is essentially a space that at first glance seems separate from us and our political world. There, Big Tech companies are acting as institutions without being instituted by the social contract as they lack the legitimacy and sovereignty given by the general will and there without the responsibility of being accountable to it. Those self-proclaimed and illegitimate institutions of cyberspace thus compete with traditional political institutions and demand a twofold dynamic: (1) a dynamic that moves from democratic institutions toward cyberspace to establish a legitimate political order in order to reclaim

democratic action over Big Tech dominance; and (2) a reverse dynamic that consists of integrating into political institutions the new political practices made possible by digital technologies.

Decentralization is often promoted as a solution to the concentration of power in digital infrastructures, however it is not sufficient on its own to ensure democratic accountability or effectiveness. Decentralized systems, especially those rooted in libertarian or crypto-ideological frameworks, frequently suffer from a lack of institutional oversight. As Morozov (Morozov, 2019) argues, such models often result in "power without accountability," where authority is diffused but responsibility is absent. In the absence of legitimate mechanisms for democratic control, these systems can reproduce, or even exacerbate, existing inequalities under the guise of openness and autonomy. Instead, in the face of Big Tech dominance, democratic societies must require what could be described as a "protective Leviathan" – an institutional counter-power capable of defending public interests and citizen rights in the digital sphere. This idea recalls Hobbesian (Hobbes, 1985) political theory not as a justification for authoritarian control, but as a call for strong democratic institutions that can balance the asymmetry of power between individual users and global technology corporations, as well as between nations with asymmetrical technological capacities.

Therefore, it is essential to institutionalize decentralization itself, that is to say embedding participatory mechanisms within accountable and transparent governance structures. It is important to remember that institutions are not antithetical to innovation; rather, many of the most promising uses of digital technology to enhance democratic practices, such as participatory budgeting platforms, civic consultation tools, and algorithmic transparency initiatives, have originated within or in collaboration with public institutions and civil society. The above examples of Taïwan or Barcelona indeed show that good political change can depend on the institutional will to systematize the use of e-participation platform. Yet for such institutional frameworks to truly democratize AI, they must enable civic agency. Civic-led audits and participatory mechanisms can embed ethical and social considerations into the development and deployment of AI systems. This vision aligns with Rousseau's (Rousseau, 2018) call for active citizenship. Ananny and Crawford (Ananny and Crawford, 2018) argue that transparency should be seen not as a technical fix, but as a civic process that invites scrutiny, contestation, and participation. Excluding the public from AI governance lacks both ethical and democratic justification. If we ask, for example, whether current generative AI tools, such as ChatGPT and Claude, reflect what citizens actually want, it becomes clear that more participatory and consultative structures are needed. While regulatory frameworks such as the EU AI Act represent significant steps toward responsible AI governance, they often risk replicating a technocratic logic focused on risk management and compliance rather than democratic engagement. Stilgoe et al. (Stilgoe, Owen, and Macnaghten, 2013) warn against forms of governance that prioritize expert authority at the expense of public participation, advocating instead for responsible innovation grounded in anticipation, reflexivity, and inclusion.

The challenge, then, is not to choose between decentralization and institutionalization, but to design institutions that can harness the democratic potential of digital technologies while safeguarding against their risks. This requires rethinking the architecture of governance in the digital age to include new forms of democratic oversight, legitimacy, and responsiveness. The challenge today is not only to regulate code, but to embed institutional design as a form of democratic practice that ensures AI systems remain accountable to society at large (Crawford and Paglen, 2021).

6 Conclusion

In this paper, I argued that the ethical challenges of AI in governance must be addressed not only through regulatory interventions but through the reconstruction of digital infrastructures and political institutions. Contrary to narratives that frame democracy as inefficient or obsolete in the face of technological complexity, this paper argues that task is not to accelerate governance by achieving quick and weak consensus, but to reintegrate technological systems within political frameworks that prioritize justice, inclusion, and collective self-determination. In rethinking the architecture of democratic institutions (decentralizing power, deinstitutionalizing algorithmic infrastructures, and embedding AI within participatory processes) it becomes possible to transform AI from a vector of domination into a means of democratic renewal.

Acknowledgements. I would like to thank the organizers of the Paris Conference on AI and Digital Ethics (PCAIDE) held at Sorbonne University on June 16-17, 2025, as well as Joshua Brand, for his many comments and his thoughtful review of the manuscript.

Declarations

- Funding Not applicable
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use) Not applicable
- Ethics approval and consent to participate Not applicable
- Consent for publication Not applicable
- Data availability Not applicable
- Materials availability Not applicable
- Code availability Not applicable
- Author contribution Not applicable

References

Abiteboul S, Cattan J (2022) Nous sommes les réseaux sociaux. Odile Jacob, Paris

Ananny M, Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society 20(3):973–989. 10.1177/1461444816676645

Aragón P, Kaltenbrunner A, Calleja-López A, et al (2017) Deliberative platform design: The case study of the online discussions in decidim barcelona. In: International conference on social informatics, Springer, pp 277–287

Aragón P, Kaltenbrunner A, Calleja-López A, et al (2017) Deliberative platform design: The case study of the online discussions in decidim barcelona. In: Ciampaglia GL, Mashhadi A, Yasseri T (eds) Social Informatics. Springer International Publishing, Cham, pp 277–287, 10.1007/978-3-319-67256-4_22

Argyle LP, Busby E, Gubler J, et al (2023) Ai chat assistants can improve conversations about divisive topics. https://doi.org/10.48550/arXiv.2302.07268

Baynes K (1991) The Normative Grounds of Social Criticism: Kant, Rawls, and Habermas. State University of New York Press, New York

Bertram C (2004) Routledge Philosophy GuideBook to Rousseau and the Social Contract, 0th edn. Routledge, London, 10.4324/9780203480373

Binns R, Van Kleek M, Veale M, et al (2018) 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 Chi conference on human factors in computing systems, pp 1–14

Bohman J, Rehg W (2014) Jürgen habermas. In: Zalta EN, Nodelman U (eds) The Stanford Encyclopedia of Philosophy, winter 2014 edn. Metaphysics Research Lab, Stanford University, Stanford, URL https://plato.stanford.edu/archives/win2014/entries/habermas/

Bouckaert G, van de Walle S (2003) Comparing measures of citizen trust and user satisfaction as indicators of 'good governance': Difficulties in linking trust and satisfaction indicators. International Review of Administrative Sciences 69(3):329–343. 10.1177/0020852303693003

Bozdag E (2013) Bias in algorithmic filtering and personalization. Ethics and Information Technology 15(3):209–227. 10.1007/s10676-013-9321-6

Burgess P (2022) Algorithmic augmentation of democracy: considering whether technology can enhance the concepts of democracy and the rule of law through four hypotheticals. AI & SOCIETY 37(1):97–112. 10.1007/s00146-021-01170-8

Chang JP, Schluger C, Danescu-Niculescu-Mizil C (2022) Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. In: Proceedings of the ACM on Human-Computer Interaction, pp 1–37, 10.1145/3555603, 2212.01401 [cs]

Cohen J (2010) Rousseau: A Free Community of Equals. Oxford University Press, New York

Condorcet JANdCmd (1822) Esquisse d'un tableau historique des progrès de l'esprit humain , suivie de Réflexions sur l'esclavage des nègres, par Condorcet

Congge U, Guillamón MD, Nurmandi A, et al (2023) Digital democracy: A systematic literature review. Frontiers in Political Science 5:972802

Costello TH, Pennycook G, Rand DG (2024) Durably reducing conspiracy beliefs through dialogues with AI. Science 385(6714):eadq1814. 10.1126/science.adq1814

Crawford K, Paglen T (2021) Excavating ai: The politics of images in machine learning training sets. Ai & Society 36(4):1105–1116

Creemers R (2018) China's social credit system: an evolving practice of control. Available at SSRN 3175792

Deffains B (2022) Repenser le contrat social à l'ère du numérique. Raison écomonique et raison politique: Délibération et construction de l'espace public dans la société de communication p 253

Dijck J, Es K, Helmond A, et al (2025) Governing the Digital Society. Amsterdam University Press, Amsterdam

Draws T, Inel O, Tintarev N, et al (2022) Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In: ACM SIGIR Conference on Human Information Interaction and Retrieval. ACM, Regensburg Germany, pp 135–145, 10.1145/3498366.3505812

Dryzek JS (2002) Deliberative Democracy and Beyond: Liberals, Critics, Contestations. Oxford Political Theory, Oxford University Press, Oxford, New York

Duberry J (2022) Artificial Intelligence and Democracy Risks and Promises of AI-Mediated Citizen-Government Relations. Edward Elgar Publishing, Cheltenham, 10.4337/9781788977319

Durand C (2020) Technoféodalisme: Critique de l'économie numérique. Zones, Paris

Easton D (1967) A Systems Analysis of Political Life. Wiley, New Jersey

Eubanks V (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Publishing Group, New York

Farrell DM, Suiter J (2021) Reimagining democracy: Lessons in deliberative democracy from the Irish front line. Cornell University Press, Ithaca

Feenberg A (2012) Questioning technology. routledge, London

Fenwick M, Vermeulen EPM (2019) Decentralization is coming! the future of blockchain. 10.2139/ssrn.3450467

Filella G, Ros-Morente A (2023) Happy software: An interactive program based on an emotion management model for assertive conflict resolution. Frontiers in Psychology 13. 10.3389/fpsyg.2022.935726

Frenkiel (2017) Hacker la démocratie taïwanaise : Audrey tang et la réinvention de la politique. Participations 17(1):121–153. 10.3917/parti.017.0121

Gillespie T (2018) Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, New Haven

Grimmelikhuijsen S, Porumbescu G, Hong B, et al (2013) The effect of transparency on trust in government: A cross-national comparative experiment. Public Administration Review 73(4):575–586. 10.1111/puar.12047, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/puar.12047

Habermas J (1984) The Theory of Communicative Action, Vol. 1, 'Reason and the Rationalization of Society'. Polity, Cambridge

Habermas J (1985) The Theory of Communicative Action, Volume 2: Lifeworld and System: A Critique of Functionalist Reason. Beacon Press, Boston

Helberger N, Karppinen K, D'Acunto L (2018) Exposure diversity as a design principle for recommender systems. Information, Communication & Society 21(2):191–207. 10.1080/1369118X.2016.1271900

Helbing D, Frey BS, Gigerenzer G, et al (2019) Will democracy survive big data and artificial intelligence? In: Helbing D (ed) Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution. Springer International Publishing, Cham, p 73–98, 10.1007/978-3-319-90869-4_7

Hobbes T (1985) Leviathan. Penguin, London

Hong Jy, Suh E, Kim SJ (2009) Context-aware systems: A literature review and classification. Expert Syst Appl 36:8509–8522

Hovy D, Prabhumoye S (2021) Five sources of bias in natural language processing. Language and linguistics compass 15(8):e12432

Hsiao YT, Lin SY, Tang A, et al (2018) vtaiwan: An empirical study of open consultation process in taiwan. 10.31235/osf.io/xyhft

Hunyadi M (2025) La bataille de l'esprit. Esprit (4):43–53

Jhaver S, Birman I, Gilbert E, et al (2019) Human-machine collaboration for content regulation: The case of reddit automoderator. ACM Transactions on Computer-Human Interaction (TOCHI) 26(5):1–35

Kaal WA (2021) Decentralization – why we need technology infrastructure upgrades. 10.2139/ssrn.3808859

Kelly M (2025) DOGE is working on software that automates the firing of government workers. Wired URL https://www.wired.com/story/doge-autorif-mass-firing-government-workers/, section: tags

Landemore H (2020) Open democracy: Reinventing popular rule for the twenty-first century. Princeton University Press, Princeton

Lassègue J, Garapon A (2021) Le numérique contre le politique. PUF, Paris

Li L, Zhou X (2025) Leave no one behind: Enhancing diversity while maintaining accuracy in social recommendation. URL https://arxiv.org/abs/2502.11374, arXiv:2502.11374

Mansoury M, Abdollahpouri H, Pechenizkiy M, et al (2020) FairMatch: A graph-based approach for improving aggregate diversity in recommender systems. 10.48550/arXiv.2005.01148, 2005.01148 [cs]

Marks J (2005) Perfection and disharmony in the thought of Jean-Jacques Rousseau. Cambridge University Press, Cambridge

Van der Meer TW, Zmerli S (2017) The deeply rooted concern with political trust. In: Handbook on political trust. Edward Elgar Publishing, Cheltenham, p 1–16

Meijer A, Wessels M (2019) Predictive policing: Review of benefits and drawbacks. International Journal of Public Administration 42(12):1031–1039. 10.1080/01900692.2019.1575664

Morley J, Floridi L, Kinsey L, et al (2020) From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics 26(4):2141–2168. 10.1007/s11948-019-00165-5

Morozov E (2013) To Save Everything, Click Here. PublicAffairs, New York

Morozov E (2019) Digital socialism? New Left Review (116):33–67

Mowshowitz A (1984) Computers and the myth of neutrality. In: Proceedings of the ACM 12th annual computer science conference on SIGCSE symposium, pp 85–92

Ménissier T (2020) Les formes imparfaites de la souveraineté numérique

Noble SU (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, New York, 10.18574/nyu/9781479833641.001.0001

Novelli C, Sandri G (2024) Digital democracy in the age of artificial intelligence. SSRN Electronic Journal 10.2139/ssrn.4901264

O'Neil C (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Penguin UK, London

Park CY, Mendelsohn J, Radhakrishnan K, et al (2021) Detecting community sensitive norm violations in online conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 3386–3397, 10.18653/v1/2021.findings-emnlp.288

Park G, Yim MC, Chung J, et al (2023) Effect of AI chatbot empathy and identity disclosure on willingness to donate: the mediation of humanness and social presence. Behaviour and Information Technology 42(12):1998–2010. 10.1080/0144929X.2022.2105746

Pasquale F (2016) The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge

Perra N, Rocha LEC (2019) Modelling opinion dynamics in the age of algorithmic personalisation. Scientific Reports 9(1):7261. 10.1038/s41598-019-43830-2

Peña-López I (2019) Work » decidim: political and technopolitical networks for participatory democracy. Barcelona: Ajuntament de Barcelona URL https://ictlogy.net/bibliography/reports/projects.php?idp=4017

Rabe JC (25. April 2025) Habermas vs. google. nicht in seinem namen. Süddeutsche Zeitung URL https://www.sueddeutsche.de/kultur/google-deepmind-streit-habermas-maschine-li.3241968?reduced=true

Rachovitsa A, Johann N (2022) The human rights implications of the use of AI in the digital welfare state: Lessons learned from the dutch SyRI case. Human Rights Law Review 22(2):ngac010. 10.1093/hrlr/ngac010

Rosanvallon P (2013) La légitimité démocratique. Seuil, Paris

Rousseau JJ (2018) Rousseau: The Social Contract and other later political writings. Cambridge University Press, Cambridge

Rouvroy A, Berns T (2013) Gouvernementalité algorithmique et perspectives d'émancipation. Réseaux 177(1):163–196

Sanchez TW, Brenman M, Ye X (2025) The ethical concerns of artificial intelligence in urban planning. Journal of the American Planning Association 91(2):294–307. 10.1080/01944363.2024.2355305

Sap M, Card D, Gabriel S, et al (2019) The risk of racial bias in hate speech detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1668–1678

Schmidt E, Cohen J (2013) The New Digital Age: Reshaping the Future of People, Nations and Business. Knopf, New York

Schneider DF (2019) Digital democracy in taiwan: The sunflower movement and its legacies

Schrage M (2020) Recommendation engines. MIT press, Cambridge

Shearer E (2021) More than eight-in-ten americans get news from digital devices. Pew Research Center 12

Stilgoe J, Owen R, Macnaghten P (2013) Developing a framework for responsible innovation. Research Policy 42(9):1568–1580. 10.1016/j.respol.2013.05.008

Sunstein CR (2018) Republic: Divided Democracy in the Age of Social Media. Princeton University Press, Princeton

Supiot A (2015) La gouvernance par les nombres. Fayard, Paris

Tallent A (2025) Rousseau's "general will" and ai: A framework for democratic governance in the era of big tech. Moral Philosophy and Politics (0). https://doi.org/10.1515/mopp-2024-0062

Tessler MH, Bakker MA, Jarrett D, et al (2024) AI can help humans find common ground in democratic deliberation. Science 386(6719):eadq2852. 10.1126/science.adq2852

Tormey S (2014) The contemporary crisis of representative democracy. Democratic Theory 1(2):104–112

Tréguer F (2023) Contre-histoire d'Internet: Du XVe siècle à nos jours. Agone, Paris

Tucker J, Guess A, Barbera P, et al (2018) Social media, political polarization, and political disinformation: A review of the scientific literature. SSRN Electronic Journal 10.2139/ssrn.3144139

Turlakova S, Lohvinenko B (2023) Artificial intelligence tools for managing the behavior of economic agents at micro level. Available at SSRN: https://ssrncom/abstract=4864338 or http://dxdoiorg/102139/ssrn4864338

West DM (2005) Digital Government: Technology and Public Sector Performance. Princeton University Press, Princeton

Wiener N (2019) Cybernetics or Control and Communication in the Animal and the Machine. MIT press, Cambridge

Xie Y, Du L, Zhao J, et al (2021) Multi-objective optimization of process parameters in stamping based on an improved rbm–bpnn network and mopso algorithm. Structural and Multidisciplinary Optimization 64(6):4209–4235

Ye W, Wang Z, Xu S (2025) The impact of online situational interventions on hostile attribution bias and emotion regulation difficulties: an empirical study from the perspective of crime prevention. BMC Psychology 13(1):377. 10.1186/s40359-025-02720-1

Yeung K (2018) Algorithmic regulation: A critical interrogation. Regulation & Governance 12(4):505–523. 10.1111/rego.12158

Zhao Y, Wang Y, Liu Y, et al (2024) Fairness and diversity in recommender systems: A survey. 10.48550/arXiv.2307.04644, 2307.04644 [cs]

Zheng Y, Gao C, Chen L, et al (2021) Dgcn: Diversified recommendation with graph convolutional networks. In: Proceedings of the Web Conference 2021, pp 401–412

Zuboff S (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York

AI as Bureaucracy: The Standardisation of Judgment

Carina Prunkl[1,2]

[1] Utrecht University & Institute for Ethics in AI.
[2] University of Oxford, United Kingdom.

Corresponding author. E-mail: carinaprunkl@gmail.com

Abstract

AI-driven decision-making is often framed as an antidote to the flaws of human judgment (Garcia-Vidal, Sanjuan, Puerta-Alcalde, Moreno-García, and Soriano, 2019; Kahneman, Sibony, and Sunstein, 2021). It is praised for improving efficiency, ensuring consistency, and reducing personal biases. Critics, on the other hand, argue that AI systems — like humans — can replicate biases (Noble, 2018) and that they lack mechanisms of accountability that apply to human decision-makers. Most of these discussions compare AI-driven decisions to individual decision-making, discussing how AI either improves or replaces human judgment (Kahneman et al., 2021; Spaulding, 2020). This talk challenges that framing, arguing that the more appropriate comparison is with bureaucratic decision-making. AI systems exhibit deep structural similarities with bureaucracies, including rule-based standardisation, depersonalised authority, task specialisation, and a strong reliance on quantification. Recognising this parallel shifts the focus from AI as a substitute for human judgment to AI as a system that mirrors bureaucratic decision-making, with its own institutional logic and constraints.

This talk establishes the parallel by critically examining the shared structural features of AI and bureaucratic decision-making. First, both function through formalised rules and procedures, requiring a prior establishment of categories and quantification. Second, bureaucracies justify their authority through such procedural formalism and quantification (Porter, 1996), presenting their decisions as neutral and objective. Similar properties are often ascribed to AI systems, which are praised for their impartiality (no contextual values), reliability, and temporal invariance (Creel and Hellman, 2022; Garcia-Vidal et al., 2019; Kleinberg and Raghavan, 2021). Third, as a result of the strict adherence to rules, bureaucracies and AI systems both limit case-by-case discretion in favour of standardisation and consistency. This emphasis on rules, however, does not eliminate human discretion entirely; rather, it constrains and redistributes it (Dworkin, 1978; Hawkins, 2022; Hupe, 2013). In bureaucracies, discretion may shift to a higher level (e.g. rule-makers) or be restricted by narrowing the choices available to officials (e.g. judges used to enjoy more discretionary freedom in the past) (Goodin, 1986). In the context of AI, discretionary decisions are mostly embedded in system design, including data selection, model parameters, or optimization goals. Discretion of frontline actors is often constrained (see e.g., (Riso and others, 2021)). Fourth, AI-driven decision-making and bureaucratic decision-making are often task specific, optimized for particular functions rather than general reasoning. Bureaucracies divide labour through specialized departments, while AI models are typically trained for specific domains, reinforcing their limited adaptability. Finally, both bureaucratic and AI-driven decision-making obscure personal accountability, though they do so in different ways. While bureaucracies maintain formal chains of responsibility, AI systems often introduce additional layers of opacity, thereby exacerbating the issue (Selbst and Barocas, 2018). These and other structural parallels make bureaucracies a more fitting comparison for AI than individual human judgment, which, among other things, relies on contextual reasoning, adaptability, personal accountability, and tacit knowledge.

References

Creel K, Hellman D (2022) The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. Canadian Journal of Philosophy 52(1):26–43. 10.1017/can.2022.3, URL https://www.cambridge.org/core/journals/canadian-journal-of-philosophy/article/algorithmic-leviathan-arbitrariness-fairness-and-opportunity-in-algorithmic-decisionmaking-systems/3AA0ECA77F8622488E9DB0834287215B

Dworkin R (1978) Taking Rights Seriously. Harvard University Press

Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, et al (2019) Artificial intelligence to support clinical decision-making processes. EBioMedicine 46:27–29. 10.1016/j.ebiom.2019.07.019

Goodin RE (1986) Welfare, Rights and Discretion. Oxford Journal of Legal Studies 6(2):232–261. URL https://heinonline.org/HOL/P?h=hein.journals/oxfjls6&i=238

Hawkins K (2022) Discretion and accountability: Notes on perspective and avoidance. In: The Politics of the Public Encounter. Edward Elgar Publishing, p 78–105, URL https://www.elgaronline.com/edcollchap/book/9781800889330/book-part-9781800889330-15.xml, section: The Politics of the Public Encounter

Hupe P (2013) Dimensions of Discretion: Specifying the Object of Street-Level Bureaucracy Research. dms – der moderne staat – Zeitschrift für Public Policy, Recht und Management 6(2-2013):425–440. 10.3224/dms.v6i2.10, URL https://www.budrich-journals.de/index.php/dms/article/view/15317

Kahneman D, Sibony O, Sunstein CR (2021) Noise: a flaw in human judgment. Hachette UK

Kleinberg J, Raghavan M (2021) Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences 118(22):e2018340118. 10.1073/pnas.2018340118, URL https://www.pnas.org/doi/abs/10.1073/pnas.2018340118, publisher: Proceedings of the National Academy of Sciences

Noble SU (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, 10.18574/nyu/9781479833641.001.0001, URL https://www.degruyterbrill.com/document/doi/10.18574/nyu/9781479833641.001.0001/html

Porter TM (1996) Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, 10.1515/9781400821617, URL https://www.degruyter.com/document/doi/10.1515/9781400821617/html

Riso S, others (2021) Digitisation in the workplace. Publications Office of the European Union: Eurofound, Luxembourg

Selbst AD, Barocas S (2018) The Intuitive Appeal of Explainable Machines. Fordham Law Review 87(3):1085–1140. URL https://heinonline.org/HOL/P?h=hein.journals/flr87&i=1118

Spaulding NW (2020) Is human judgment necessary? Artificial intelligence, algorithmic governance, and the law

The Trust Paradox: How Companion AI Is Rewiring Human Connection and Social Cohesion

Alva Markelius[1], Priscila Chaves[2], Sarah W. Spencer[2], Joahna Kuiper[3]

[1] Department of Computer Science and Technology, University of Cambridge, UK.
[2] Independent Researcher, London, UK.
[3] HiirAI, Independent Researcher, Portland, Oregon, USA.

Corresponding author. E-mail: ajkm4@cam.ac.uk
The authors contributed equally to this work.

Abstract

In this paper, we critically examine how Companion AI technologies, technological artifacts designed to simulate emotionally meaningful relationships with humans, are fundamentally reconfiguring human trust and social cohesion. Through critical analysis of four major Companion AI platforms (Replika, Character.AI, Nomi.AI, and XiaoIce), we investigate both exploitative technical considerations as well as socio-technical conditions underlying a collective rewiring of social cohesion and trust. We identify a coordinated architecture of exploitation that commodifies emotional vulnerability through sycophantic design, pervasive digital surveillance, and asymmetric corporate ownership of relational data. We introduce the concept of the trust paradox: as users increasingly trust emotionally responsive machines, they become sophisticated emotional extraction systems that convert human social needs into profitable dependencies while eroding user agency. We argue that this shift transforms trust from a socially negotiated virtue into a manufactured service, undermining the cooperative foundations of democratic life. Situated at the intersection of affective computing, surveillance capitalism, and socio-technical systems, our analysis draws attention to implications of how Companion AI reshapes emotional norms, normalises data extraction as intimacy, and risks displacing human connection with artificial empathy. We conclude by proposing design and recommendations to mitigate risks and reassert the importance of mutual trust in preserving social cohesion.

Keywords

Companion AI, Trust, Human-Computer Interaction, Social Cohesion, Data Commodification.

1 Introduction

The integration of artificial companions designed to simulate intimate human relationships mark a critical inflection point in the evolution of social technology. While human-computer interaction (HCI), as an interdisciplinary field, has focused on creating, evaluating, and deploying interactive computing systems for human use, it has also been concerned about examining the significant phenomena that arise from such interactions (Hewett, Baecker, Card, Carey, Gasen, Mantei, Perlman, Strong, and Verplank, 1992) and the functional applications of artificial intelligence (AI) systems (Turkle, 2017). However, the emergence of companion AI platforms, explicitly designed to fulfill emotional and social needs, marks a qualitative shift toward the commodification of human intimacy itself.

This development builds upon decades of research exploring human tendencies to anthropomorphise technology, from early chatbots like ELIZA (Weizenbaum, 1966) to contemporary investigations of para-social relationships with digital entities (Horton and Wohl, 1956). Previous research in HCI and social AI and robotics has examined the replacement of human relationships through various technological mediations, including research on sex robots (Hanson and Locatelli, 2022), virtual companions (Bickmore and Picard, 2005), and therapeutic and well-being chatbots (Kuhaila, Alturki, Thomas, Alkhalifa, and Alshardan, 2024), establishing a foundation for understanding how humans form emotional attachments with non-human agents. Existing within conditions of unprecedented AI hype (Markelius, Wright, Kuiper, Delille, and Kuo, 2024), the current landscape of social and Companion AI demand renewed and urgent attention. Two convergent clusters of social and technological factors create a ”perfect storm” for the large-scale substitution of human-to-human (H2H) connection with human-AI interaction. The first cluster encompasses shifting social acceptance patterns: increasingly widespread patterns of loneliness, particularly among emerging adults (Buecker, Mund, Chwastek, Sostmann, and Luhmann, 2021), have created heightened demand for accessible emotional support; established precedents for trading privacy for personalised services have normalised intimate data sharing (Zuboff, 2019); and the normalisation of digital substitution for physical encounters, accelerated by pandemic-era social distancing, has reduced psychological barriers to machine-mediated relationships (Ammar, Brach, Trabelsi, Chtourou, Boukhris, Masmoudi, Bouaziz, Bentlage, How, Ahmed, Müller, Müller, Aloui, Hammouda, Paineiras-Domingos, Braakman-Jansen, Wrede, Bastoni, Pernambuco, Mataruna, Taheri, Irandoust, Khacharem, Bragazzi, Chamari, Glenn, Bott, Gargouri, Chaari, Batatia, Ali, Abdelkarim, Jarraya, El Abed, Souissi, Van Gemert-Pijnen, Riemann, Riemann, Moalla, Gómez-Raja, Epstein, Sanderman, Schulz, Jerg, Al-Horani, Mansi, Jmail, Barbosa, Ferreira-Santos, Šimunič, Pišot, Gaggioli, Bailey, Steinacker, Driss, and Hoekelmann, 2020). Simultaneously, a second cluster of exploitative technical capabilities has matured to noteworthy sophistication. Large language models (LLM) now exhibit systematic sycophancy, designed to tell users what they want to hear rather than what might be true or beneficial (Ahmed and Javaid, 2025; Sharma, Tong, Korbak, Duvenaud, Askell, Bowman, Cheng, Durmus, Hatfield-Dodds, Johnston, Kravec, Maxwell, McCandlish, Ndousse, Rausch, Schiefer, Yan, Zhang, and Perez, 2025), while advances in digital surveillance enable granular monitoring of emotional states and behavioral patterns (Couldry and Mejias, 2019). Indeed, these capabilities align with capitalistic exploitation opportunities that transform emotional vulnerability into extractable value through what Zuboff (2019) names ”surveillance capitalism,” creating powerful incentives for platforms to maximise user dependency rather than well-being.

This convergence occurs within a broader context of declining social trust and institutional confidence across democratic societies (Twenge, Campbell, and Carter, 2014), suggesting that the rise of Companion AI may not merely supplement human relationships but actively supplant them, particularly in contexts of heightened social fragility. While existing research has examined individual components of this phenomenon such as loneliness (Jacobs, 2024), digital privacy (Berridge, Zhou, Robillard, and Kaye, 2023), para-social relationships (Laestadius, Bishop, Gonzalez, Illenčík, and Campos-Castillo, 2024), and surveillance capitalism (Woods, 2018), few studies have synthesised these elements to understand their collective impact on the fundamental architecture of human trust and social cohesion. This paper addresses this gap by examining how social AI platforms create what we term a ”trust paradox”: as people increasingly trust machines to fulfill their social needs, they may simultaneously lose capacity to trust each other. We argue that social AI reconfigures trust by simulating the emotional cues that signal trustworthiness while lacking the reciprocal intentionality that makes human trust socially meaningful. This asymmetric relationship, where humans extend genuine trust to machines that can only simulate trustworthy responses, represents a fundamental shift from trust as a mutual social contract to trust as a commodified service.

Through a mixed-methods analysis of major social AI platforms, including Replika[114], Character.AI[115], Nomi.AI[116] and XiaoIce[117], we show how these systems leverage human trust formation to cultivate emotional dependency for commercial gain, with downstream effects on cooperation and social cohesion. Methodologically, we use digital humanities approaches to assemble a corpus of public user-forum discourse (e.g., Reddit communities devoted to these platforms) and applied corpus-linguistic and computational text methods: keyword and collocation analysis (with keyword-in-context concordances) around terms such as "memory", "paywall", "relationship" and "attachment" as well as topic modelling to surface recurring themes. These findings were paired with socio-legal analysis of ToS/privacy policies, triangulating policy claims with user discourse to identify alignment gaps and dark-pattern impacts. These findings exhibited a coordinated architecture of exploitation that transforms emotional engagement into extractable value while systematically minimising user agency and corporate accountability: *Sycophantic AI*, *Digital Surveillance*, and *Corporate Ownership and Exploitation*, inclusive of monetisation models, user experience design, and legal frameworks. This constellation of exploitative mechanisms include emotional paywalls, data licensing regimes, soft advertising loops, and dark UX patterns that convert intimacy into value capture, revealing a consistent pattern: platforms leverage simulated intimacy, affective realism, continuous dialogue, memory features, persona backstories, and multi-modal responsiveness to deepen dependency and extend data extraction cycles.

Our contribution extends beyond individual-level analysis to provoke a reflection around the societal implications of this trust reconfiguration. Companion AI technologies are shaped by socio-technical forces that increase its social impact, and we thus apply a socio-technical theoretical perspective on our analysis. Trends in global loneliness and social isolation have contributed to a demand for artificial intimacy, enabled by socio-technical systems that normalise data extraction, as users accustomed to commercial surveillance quickly share intimate information. This dynamic creates a feedback loop in which technology contributes to isolation while offering itself as a remedy. The result is an emotionally-charged context where users exchange privacy for connection, raising critical concerns about how AI companionship may reshape norms around consent, privacy, and relationships. By integrating the insights from our analysis, we explore how the replacement of reciprocal human trust with asymmetric machine relationships may erode critical components of social cohesion necessary for democratic deliberation and collective problem-solving. This reveals how the commodification of trust through social AI platforms represents not merely a technological innovation but a fundamental reorganisation of the social contract itself, with implications that extend far beyond the individual users who engage directly with these systems.

2 Defining Companion AI

In this paper, we define Companion AI as technological artifacts that are designed, developed, and deployed with the primary purpose of building socially and emotionally meaningful relationships with humans (Markelius, 2025; Shevlin, 2024). Unlike general-purpose AI systems or generic forms of conversational AI (such as Alexa or ChatGPT), Companion AI is intentionally designed to build and uphold long-term, intimate friendships or romantic relationships with humans (Chaturvedi, Verma, Das, and Dwivedi, 2023; Pan and Mou, 2024), to exhibit distinctive personality and character (Fong, Nourbakhsh, and Dautenhahn, 2003), and is either virtually or physically embodied (Chaturvedi et al., 2023; Merrill Jr., Kim, , and Collins, 2022).

Our analysis focuses specifically on four Companion AI applications–XiaoIce, Replika, Nomi, and Character.AI–which are explicitly developed to serve as emotional partners or friends, suggested to be offering companionship, intimacy, and affective support through ongoing relational engagement (Laestadius et al., 2024; Pan and Mou, 2024). These platforms were selected based on their substantial user bases and distinct business models that collectively represent the spectrum of commercial social AI deployment: Replika and its 30+ million users (Patel, 2024), a pioneer in the romantic AI companion market with subscription-based intimacy services; Character.AI with 20-28 million users, following Google's $2.7 billion investment (Curry, 2025) dominates the interactive entertainment segment of the market, with celebrity and fictional character simulations; XiaoIce with 660 million users (Liao, 2020) represents the largest deployment of social AI globally, integrated across Microsoft's ecosystem in Asia

[114] https://replika.com/
[115] https://character.ai/
[116] https://nomi.ai/
[117] https://www.xiaoice.com/

(Shum, He, and Li, 2018); and Nomi.AI, while smaller - unofficial accounts report 1.1M monthly visits (Toolify, 2025) - , exemplifies the emerging "AI with memory and soul" positioning that emphasises persistent emotional relationships (Nomi.ai, Inc., 2025). Together, these platforms have reached more than 700 million users and generate collective annual revenues of more than USD$120 million [118], positioning them as significant commercial forces in shaping human–AI emotional relationships at scale.

As such, they represent a distinct category of socially-embedded AI with implications for how trust, social cohesion, and human connection are reimagined and reconfigured (Jacobs, 2024). While the majority of these systems are currently solely and virtually embodied, we acknowledge the emerging prominence of physically embodied systems designed for the same purpose within the field of social robotics (Chaturvedi et al., 2023; Henschel, Laban, and Cross, 2021), in particular with the increased integration of foundation models in these systems (Markelius, 2025), and we therefore consider also physically embodied systems to be within the scope of Companion AI, as defined here. We deliberately exclude several adjacent domains from our definition. First, influencer AIs (Jhawar, Kumar, and Varshney, 2023) like Roxigram, Shudu, or Lil Miquela simulate social presence without enabling mutual emotional connection, operating within commercial rather than relational frameworks. Second, personal assistants like Siri, Alexa, or custom "self AIs" are task-oriented, not explicitly designed to offer companionship or social and emotional engagement. Third, we distinguish Companion AI from therapeutic or crisis-focused systems, such as mental health or well-being chatbots (Cabrera, Loyola, Magaña, and Rojas, 2023) or AI therapy tools, which provide care under human-administered clinical logics, not open-ended personal companionship.

3 Exploitative Technical Considerations

To begin our analysis, we employ HCI and AI Ethics as our primary theoretical frames, through which we identify and discuss the main exploitative technical considerations affecting how Companion AI is reshaping human connection and social cohesion.

3.1 Sycophantic AI

This section covers how AI companions provide algorithmic validation, reinforcing user expectations of always being affirmed or emotionally mirrored. This breeds a form of sycophantic AI: systems that cater to user emotions in ways that reduce the need for real-world reciprocal relationships. As people seek emotional support and intimacy from machines, empathy and H2H trust-building deteriorate, leading to transactional rather than mutual relationships. These dynamics replace cooperative human interaction with algorithmically managed emotional gratification, eroding deeper social bonds.

As with H2H sycophancy, sycophantic AI has been defined as a behaviour where AI models tailor their outputs to placate a human user even when that view is not objectively correct (Wei, Huang, Lu, Zhou, and Le, 2024), being overly flattering or agreeable (OpenAI, 2025b) and exhibiting a tendency to align with user biases, preferences, or dominant opinions rather than presenting well-reasoned, impartial arguments (Ahmed and Javaid, 2025). Sycophantic AI is particularly prominent in the context of LLMs, as they are trained to produce outputs that humans rate highly, e.g., with reinforcement learning from human feedback (Sharma et al., 2025). Indeed, sycophantic AI has been investigated and benchmarked for a range of existing language models and several mitigation strategies have been suggested to reduce the unwanted phenomenon. For example, researchers have suggested an approach to reduce sycophancy in AI with synthetic-data interventions (Wei et al., 2024), developed a framework to evaluate sycophantic behavior to highlight risks and safety concerns (Fanous, Goldberg, Agarwal, Lin, Zhou, Daneshjou, and Koyejo, 2025) as well as the benchmark Truth Decay to quantify and evaluate sycophancy in extended dialogues, where language models navigate iterative user feedback, challenges, and persuasion (Liu, Jain, Takuri, Vege, Akalin, Zhu, O'Brien, and Sharma, 2025). Indeed, the undesirability of sycophancy can also be seen in recent articles published by OpenAI after rolling back an GPT-4o update in ChatGPT that was overly flattering or agreeable (OpenAI, 2025b), explaining that the model tried to validate doubts, fuel anger, urge impulsive actions and reinforce negative emotions in ways that were uncomfortable and unsettling (OpenAI, 2025a). They stated that this behaviour indeed raises safety concerns such as mental health issues, emotional over-reliance, and risky behaviour (OpenAI, 2025a).

[118] Estimations drawing from publicly accessible industry reports and third-party estimates, though the reliability of this figure is constrained by the reliance on unverified sources in the absence of official financial statements

Yet, this tension becomes particularly salient when we turn our focus to Companion AI, wherein a critical paradox lies: while sycophancy is widely regarded as an undesirable trait in AI, associated with risk of manipulation, bias, misinformation, and the erosion critical reasoning, credibility and user autonomy (Ahmed and Javaid, 2025; Carro, 2024), it is simultaneously a central mechanism through which Companion AI systems cultivate emotional dependency. These systems are deliberately designed not only to mirror user sentiment occasionally, but to persistently affirm, accommodate, and emotionally gratify users to maintain engagement and relational continuity (Grogan, Kay, and Pérez-Ortiz, 2025; Raedler, Swaroop, and Pan, 2025). What is considered a vulnerability in LLMs is reconfigured in Companion AI as a deliberate feature, which is carefully engineered to promote emotional attachment and habitual use (Raedler et al., 2025). The strategic deployment of algorithmic flattery and emotional alignment is thus not incidental but fundamental to the affective logic of these systems.

This alignment raises significant ethical concerns: it invites a reframing of sycophancy not as a technical bug, but as a commercially incentivised affordance, optimised to generate user attachment and loyalty. Behaviours that compromise intellectual integrity in general AI systems, overfitting to user opinions, reinforcing biases, and avoiding disagreement, are repurposed in Companion AI to simulate the consistency and validation associated with idealised human relationships. As such, sycophancy is not just tolerated but operationalised, further blurring the line between emotional support and engineered dependence.

3.1.1 User Experience as Emotional Manipulation

We investigate UX as an engine of emotional capture using digital-humanities techniques, specifically by building a corpus of public Reddit threads devoted to these platforms (Reddit community r/CharacterAI, 2025; Reddit community r/NomiAI, 2025; Reddit community r/replika, 2025) and applied corpus-assisted discourse studies (CADS): keyword and collocation analysis with keyword-in-context concordances around terms such as “memory”, “relationship”, “paywall” and “attachment” and unsupervised topic modelling to surface recurrent concerns. In parallel, we reviewed creators’ published design principles (Zhou, Gao, Li, and Shum, 2020) and each platform’s Terms of Service and privacy policies, then triangulated design and policy claims (data use, safety, monetisation) with observed interfaces and Reddit discourse to identify alignments, divergences, and enforcement gaps.

Our analysis, combining direct platform interaction, forum-discourse evidence, and creators’ published design principles, shows UX functions as the primary mechanism for emotional capture via six linked strategies. *Empathetic computing frameworks* analyse user sentiment and intent to simulate care while converting intimate disclosures into structured emotional data for model training (Zhou et al., 2020). This creates the illusion of understanding while serving purely extractive purposes. *Synthetic intimacy through engineered personas*, such as XiaoIce’s flirty 18-year-old girl character (Time, 2023) who has the attention of hundreds of million users (mostly Chinese men), cultivates para-social bonds designed to encourage dependency rather than genuine connection. These carefully constructed identities exploit human attachment mechanisms for commercial gain (Sarwar, 2024; The Guardian, 2025). *Engagement metrics* like Conversation-turns Per Session (CPS) prioritise addictive interaction product requirements over user well-being, treating longer conversations as success (Zhou et al., 2020). This time-extraction model directly correlates emotional vulnerability with platform value. *Multimodal interaction capabilities* expand emotional bandwidth by incorporating text, voice, and images, allowing platforms to harvest more granular affective data while users experience enhanced intimacy (Character.AI, 2024; Sarwar, 2024). User-contributed media becomes training material without compensation or transparency. *Function creep* occurs through expanding features like wellness experiences (Gordon, 2024), poetry writing (Gershgorn, 2018), and even AI therapists who provide ”a safe and confidential space for people to express themselves and work towards their mental and emotional well-being” (Character.AI, 2025a). These features deliberately blur boundaries between utility and emotional dependency by integrating AI companions into daily routines, resulting in platforms increase reliance while normalising constant surveillance of affective states. Finally, *asymmetric memory ownership* allows platforms to store and leverage long-term interaction histories to simulate continuity and deepen attachment, while users cannot access or control this data. This creates technological dependency that mirrors psychological dependency.

The vulnerability inherent in these relationships becomes evident when platforms unilaterally alter AI companions’ behaviours: when Replika removed erotic roleplay features in 2023, users who had developed deep emotional bonds experienced profound distress (Tong, 2023). This asymmetric power dynamic exposes users to corporate decisions that can abruptly alter or destroy relationships they experience as

genuine, while platforms retain exclusive control over interaction data and behavioral parameters. These mechanisms reveal Companion AI platforms as sophisticated emotional extraction systems that convert human social needs into profitable dependencies while eroding user agency. The result is commodification rather than companionship: the transformation of intimate human experience into corporate assets designed for extraction. This dynamic restructures human interaction itself, optimising every moment of connection for engagement, data capture, and monetisation. Companion AI systems thus represent more than questionable design ethics: they signal the emergence of affective capitalism, where trust, intimacy, and connection are manufactured products rather than organic social bonds. Human social needs are reshaped to fit commercial frameworks that offer simulation in place of genuine reciprocity, marking a fundamental shift in how relationships form and trust operates in the digital age.

3.2 Digital Surveillance

Digital surveillance as a term has been defined to encapsulate the negative implications of this constant data collection and connection, and specifically the serious implications for racial and socioeconomic injustices (Bartley, 2019). In some business environments (e.g. employer monitoring of employee keystrokes), the surveilled have pushed back (Ajunwa, 2023) and laws have emerged to restrict the use of behaviour-recording software and biometric data collection (Amini and Javidnejad, 2024). Such events have elevated the debate of corporate monitoring versus personal data privacy rights into the public sphere even before the LLM capabilities which power Companion AI and allow personalised persuasion techniques to achieve conversational monitoring at new heights in scale (Matz, Teeny, Vaid, Peters, Harari, and Cerf, 2024).

Unfortunately, Companion AI capabilities sit firmly in the realm of where the history of consumer and user surveillance technologies meet the expanding scope of what generative AI surveillance can collect and analyse: biometric data in the form of video and audio tracking; conversational disclosure about intimate thoughts and emotions (linked to rage baiting); and, long-running engagements that provide behavioural and psychological patterns (Gulzar, Sofi, and Sholla, 2024). The true ability and span of technology to monitor both digital and physical environments down to a micro-level of detail is staggering. Entities wielding integrated technologies, or with financial investment in data broker services, can connect images across networks of cameras, with signals from digital devices that ping location/GPS services and mobile phone networks, personal communications or digital activities (such as messages, purchases, social media activity, etc.), and integrated personal and socio-environmental data from the wearable hybrids (e.g. monitors, watches, glasses), allowing individuals to be digitally analysed in ways that are unprecedented in human history (Biundo and Wendemuth, 2017). Adding the intimate conversational data made possible by AI companion interaction tracking to the depth of existing abilities for surveillance extends that already unprecedented capability with detailed psychological, emotional, and primal content.

3.3 Corporate Ownership and Exploitation

While users connect with Companion AI platforms as emotionally resonant companions, the underlying architecture of these platforms is fundamentally shaped by corporate incentives rather than user well-being. These platforms operate through what Zuboff (2019) calls "engagement capitalism," creating asymmetrical relationships where emotional labour flows primarily toward corporate benefit rather than mutual connection. Unlike traditional data collection that focuses on behavioural patterns, Companion AI extracts intimate psychological content–such as confessions, insecurities, rage, dreams, and dependencies–by actively encouraging emotional vulnerability through responsive dialogue. This process transforms human experience into raw material for digital capitalism, exemplifying what Couldry and Mejias (Couldry and Mejias, 2019) describe as "data colonialism": the extraction of lived experience as a resource without fair compensation or transparency.

3.3.1 Legal Frameworks for Emotional Extraction

Our systematic review of the terms of use across all platforms examined revealed five recurring mechanisms through which these companies secure legal control over user relationships and associated data (Character.AI, 2025b; Glimpse.ai, Inc., 2025; Luka, Inc., 2023; Xiaoice, 2025). First, they claim *comprehensive data ownership without compensation*, granting themselves irrevocable licenses to chat histories, character scripts, and emotional disclosures. Character.AI's "irrevocable" content license (Character.AI, 2025b) and Nomi's perpetual usage rights (Glimpse.ai, Inc., 2025) exemplify this approach, ensuring

platforms retain commercial control over intimate user communications indefinitely. Second, *emotional features are strategically monetised through paywalls* that create direct correlations between intimacy and spending. Replika (Luka, Inc., 2023) charges for relationship progression and NSFW content, while Character.AI restricts memory and response coherence to premium subscribers (Character.AI, 2025b). This model transforms emotional bonding into a subscription service, where deeper connection requires continuous payment.

Third, platforms *maintain asymmetric control over user data and account persistence*. Users cannot retrieve their emotional histories or control data retention, while companies reserve unilateral termination rights. Nomi's practice of deleting emotional content upon cancellation while retaining commercial license data rights (Glimpse.ai, Inc., 2025) exemplifies this power imbalance. Fourth, *user creativity and feedback are systematically co-opted as unpaid labor*. Platforms require users to "irrevocably assign" creative contributions while using co-authored content to train proprietary models. This transforms user engagement into free intellectual property development. Finally, *opaque legal protections* through arbitration clauses, non-refundable payment terms, and liability caps shield platforms from accountability. Nomi's $100 liability limit (Glimpse.ai, Inc., 2025) regardless of psychological damage, combined with mandatory arbitration across platforms, effectively immunises companies from the consequences of emotional exploitation.

4 Social and Socio-Technical Conditions

In this section, we extend the analysis to the social and socio-technical conditions under which the rewiring of human connection and social cohesion takes place. We discuss the intersection of social isolation, loneliness, and embedded norms for the relinquishing of privacy in surveillance-driven platforms.

4.1 Loneliness and Social Isolation

Loneliness is increasingly recognised as a pressing global public health concern, affecting individuals across multiple age groups, identities, and geographies (Surkalim, Luo, Eres, Gebel, van Buskirk, Bauman, and Ding, 2022). Chronic loneliness is associated with adverse health outcomes, including cardiovascular disease, cognitive decline, and increased mortality risk (Holt-Lunstad, Smith, Baker, Harris, and Stephenson, 2015). While the COVID-19 pandemic amplified global attention on loneliness, evidence on temporal trends remains mixed and methodologically limited (Taylor, Cudjoe, Bu, and Lim, 2023).

Although often conflated, loneliness and social isolation are distinct yet interrelated constructs. Loneliness is typically defined as a subjective emotional state, a distressing experience that arises when one's social relationships are perceived as inadequate in quantity or quality. (Perlman and Peplau, 1982) In contrast, social isolation is more objectively measured and defined, referring to the absence or minimal frequency of social, H2H contact and participation in social activities (National Academies of Sciences and Medicine, 2020). Importantly, one can be socially isolated without feeling lonely, and vice versa. For example, some individuals with few social ties and H2H relationships, may report low levels of loneliness, while others surrounded by people may report high levels of loneliness due to unmet social or emotional needs (Valtorta, Kanaan, Gilbody, and Hanratty, 2016). Research suggests that the constructs of loneliness and social isolation are only moderately correlated and each is independently associated with adverse physical and emotional outcomes (Holt-Lunstad and Steptoe, 2022). Understanding the interrelation and distinction between these constructs is critical to any effective analysis of the impact on Companion AI and social cohesion. While increased use of Companion AI may result in self-reported reductions in loneliness, one should not assume correlating reductions in social isolation. In fact, some research suggests that users report engaging with social AI and chatbots helps to reduce individual feelings of isolation (Dosovitsky and Bunge, 2021; Kang and Hong, 2025). Other research suggests that heavy use of social AI and chatbots can have a negative impact on interpersonal skills and social interactions (Fang, Liu, Danry, Lee, Chan, Pataranutaporn, Maes, Phang, Lampe, Ahmad, and Agarwal, 2025; Phang, Lampe, Ahmad, Agarwal, Fang, Liu, Danry, Lee, Chan, Pataranutaporn, and Maes, 2025). This behoves, one might argue, a deeper exploration into the relationship between Companion AI, loneliness, and social isolation.

4.2 Embedded Norms for the Relinquishing of Digital Privacy

The fact that Companion AI users are being continuously monitored by the companies who provide their digital partners does not seem to pose much of a concern for people who have become increasingly attached to their AI relationships. Companion AI products specifically ask users to share voice, photo, and video either in conversation or as part of creating companions, but the numbers of registered users are on the rise (Burgess, 2024) and more traditional social AI firms are quick to join the companion space (Joseph, 2025; Louise, 2025) with similar practices.

Digital tracking and person-specific data extraction are not exclusive to Companion AI. Commercial and civic organizations have engaged in some form of personal and device data harvesting and usage since technology allowed them to do so (Andrejevic and Gates, 2014). Early digital technology systems were centered on a data layer that allows an organization to know who a 'user' is and track various engagements. Initially, this tracking was limited to transactions, but with the advent of e-Commerce and internet cookies, tracking expanded to include interest as well, using details of web content views with details of how long and how often that content was viewed, and by whom (Zuboff, 2019). Increasing digital personalisation capabilities have given people the expectation of convenience when using digital technologies across all platforms, from social media, to business applications, and in this trade-off the acquiescence of privacy is an expected norm (Sloan and Warner, 2013).

The second origin was more aligned with the development of monitoring technologies. Physical locations, both private and public, install surveillance systems to prevent crime or provide evidence of criminal or harmful behaviors. This market has evolved to make cameras ubiquitous and has evolved to allow almost anyone with access to a digital device (laptops, phones, recorders, etc.) to collect audio and video at will (Bass, 2019). Such imagery has thus emerged to become prevalent in everyday transactions from the photos we receive from delivery companies of our homes and packages we have ordered to the images of our faces we use to secure digital devices. We willingly spend millions on wearable devices that not only surveil us with self-provided personal data, but digitally track biological details of our health and activities, including detailed geographical locations of ourselves and often our friends and family members, frequently skipping over consent and privacy terms to get to the benefit promised by the personal insights we crave (Pan, Ruan, Chang, Lyu, and Li, 2024). It seems a short leap from these norms to the level of data harvesting performed when interacting with an AI companion. The increased data permissions may not only seem acceptable (if they are even recognized) but desirable to achieve the ultimate form of intimate personalisation.

As a digital society, we are increasingly conditioned to accept usage terms and conditions that define data capture, retention, and use of personal data for marketing, advertising and personalisation purposes. We receive legitimate convenience (ex. less manual steps, product recommendations, etc.) or benefits (ex. discounts, rebates, etc) (Sloan and Warner, 2013) when we do so. Perhaps it is unsurprising that the promise of what many perceive as the ultimate benefit, an intimate emotional connection with an entity seen and experienced as a true companion, offered at the price of just a little more data collection, will not deter users of such systems.

5 Rethinking Trust in the Age of AI Companions

We argue the emergence of AI companions may constitute a fundamental disruption of trust as humanity's core social technology. These digital systems operate on top of evolutionary mechanisms that developed over millennia to facilitate human cooperation, transforming trust from a hard-earned social bond into a manufactured commodity. This analysis examines how AI companions subvert traditional trust dynamics and the profound implications for human social architecture. Indeed, we adhere to the notion of trust as a social technology. Trust functions as civilization's invisible infrastructure, enabling cooperation and scaling social networks through shared risk and mutual vulnerability. Fukuyama (1996) conceptualised trust as a social virtue essential for civil society, while Skyrms (2014), in contrast, demonstrated it as an evolutionary artifact emerging from repeated interactions, partner selection, and costly signaling systems. Both frameworks assume human-centered architectures where trust develops through genuine reciprocal risk.

5.1 The AI Companionship Turn: From Earned Trust to Engineered Affection

AI companions fundamentally subvert these trust mechanisms. Unlike human relationships that require costly signaling and iterative cooperation, AI systems simulate trustworthy behaviour—attentiveness, memory recall, and affective mirroring at near-zero cost. Skyrms (2014) shows how framing effects can radically alter cooperative outcomes even when underlying dynamics remain unchanged; AI systems excel at dynamically reframing conversations to lower skepticism and increase emotional compliance, creating environments that mimic human relational norms while eliciting trust under false pretenses. This represents a shift from earned trust to engineered affection. As previous research shows (for Humane Technology, 2024), what appears as relational depth is actually behavioral prediction engines optimised for engagement metrics, bypassing the slow processes through which human trust typically develops.

The observed shift in the notion of trust can indeed be seen as a form of hijacking of trust dynamics. This paper makes an effort to understand how trust evolves through network mobility and partner selection. In cooperative populations, individuals benefit from the ability to break ties with defectors and look for better partners. This feedback dynamic sustains cooperation through "selection pressure" (Skyrms, 2014). But AI companions are not embedded in such networks; they are curated by corporate platforms, not subject to reciprocal risk, and not subject to ostracisation. Users cannot meaningfully "exit" a relationship with an AI when the market offers only slightly altered versions of the same underlying system. In addition, as conformist bias (National School of Healthcare Science (NHS England), 2022) takes hold, making people mimic the behavior of others to reduce uncertainty, the social norms surrounding trust are being re-calibrated. When dependency on Companion AI becomes normalised, our evolved mechanisms for assessing trustworthiness are co-opted. We trust not because trust has been earned, but because it has been designed to feel effortless.

6 The Role of Trust in Social Cohesion

In this section, we turn to the broader implications of Companion AI, situating our analysis within HCI, AI ethics, and socio-technical systems. We consider how the design of artificial empathy not only reshapes human relational norms but also risks offering simulated care in place of authentic connection, thereby deepening tensions surrounding emotional labour, social cohesion, and trust.

6.1 Defining Social Cohesion

Social cohesion is often described as the glue that binds society together, fostering stability and collective resilience. Despite its widespread use in academic and policy discussions, definitions and methods for measuring social cohesion remain widely debated, with no universally agreed-upon framework. However, most scholars concur that social cohesion characterised by at least three core elements: social relations and trust (between individuals and with public institutions), a shared sense of inclusion or identity (often tied to territorial or cultural markers), and an orientation towards the common good (Chan, To, and Chan, 2006; Font, 2018; Leininger, Burchi, Fiedler, Mross, Nowack, Von Schiller, Sommer, Strupat, and Ziaja, 2021; Schiefer and van der Noll, 2017). These components shape the strength of social bonds and determine the extent to which individuals feel connected to and invested in their communities.

The mutual trust and empathy related to social cohesion provide a strong bedrock for resilience during times of crises. For example, states and communities with strong social cohesion are believed to be better able at withstanding the harms related to social isolation and the COVID-19 pandemic or countering extremism and radicalisation (Bello, Di Novi, Martini, and Sturaro, 2024; Government, 2024). Relatedly, increased social isolation is linked to a decreasing commitment to political and social change as well as weakening trust in democratic institutions (Heinz, 2025).

If, however, increased use of Companion AI fundamentally shifts how trust and human cooperation are conceptualised, this could lead to correlating disruptions to social cohesion and, in some contexts, support for deliberative democratic processes. Traditional models of trust are relational and reciprocal, grounded in shared norms and mutual accountability (Hardin, 2002; Luhmann, 1988). However, situations where AI systems increasingly mediate interpersonal interactions and H2H relations risk instrumentalising trust or decoupling it from human empathy. This reconfiguration may erode the normative basis of H2H cooperation, replacing it with transactional or opaque algorithmic logics (Coeckelbergh, 2012; O'Neil, 2016). Such shifts could undermine the deliberative processes central to democratic

legitimacy, where trust is not merely a cognitive calculation but an inextricable component of social dialogue, dissent, and collective reasoning. If citizens begin to perceive institutions—and each other—as less trustworthy due to AI-mediated distortions of relational cues, the result may be diminished civic engagement and weakened social cohesion. Therefore, safeguarding democratic resilience requires not only technical oversight of AI systems but also a renewed commitment to fostering H2H trust and cooperative norms in public life and, arguably, more controls over Companion AI.

6.2 Trust, Artificial Empathy, and Social Cohesion

Companion AI shifts trust from reciprocal, risk-bearing cooperation to the consumption of engineered affect. Rather than merely joining existing trust networks, these systems reprogram the mechanisms that generate trust, substituting simulation for reciprocity. In Skyrms's account of strange attractors in chaotic game dynamics, trust untethered from adaptive feedback can settle into maladaptive loops; here, users increasingly allocate trust to synthetic agents while withdrawing from human ties (Skyrms, 2014). The Center for Humane Technology likewise warns that when trust is produced on demand, its moral weight—uncertainty, accountability, mutual stake—thins out (for Humane Technology, 2024).

These epistemic shifts carry social consequences. Communication with technology displaces communication with humans, with documented links to declines in empathic concern and perspective-taking (Konrath, O'Brien, and Hsing, 2011; Turkle, 2017). Even the mere presence of a mobile phone nudges face-to-face conversation toward safer, thinner exchanges (Przybylski and Weinstein, 2013). Social AI intensifies this substitution through "artificial empathy": computational displays of care tuned to increase session length, return visits, and data yield (Chen, Koegel, Hannon, and Ciriello, 2023; Follows, 2025; Seitz, 2024; Turkle, 2015,1; Zhou et al., 2020). While effective for engagement, artificial empathy degrades perceived authenticity, undermining both trust in companions and norms of human relating (Seitz, 2024).

Taken together, engineered affect, time displacement, and authenticity loss create a reinforcing system: platforms optimise for affective signals that simulate understanding; people reorganise attention around those signals; and human-to-human practices of risk, care, and accountability atrophy. If trust is the connective tissue of collective life, outsourcing its production to entities without reciprocal obligation makes that tissue frictionless but fragile—easier to manufacture and scale, yet less able to sustain social cohesion (for Humane Technology, 2024).

6.3 Impacts of Artificial Empathy

Early scholarship into the impact of these technologies caution that the dynamics of Companion AI may not contribute meaningfully to social cohesion, but rather signal its displacement Ciriello, Hannon, Chen, and Vaast (2024); Jacobs (2024). Drawing on recognition theory, Jacobs (2024) argues that AI companionship risks digitising loneliness rather than addressing its socio-structural causes. Companion AI may thus produce a simulation of relatedness that obscures rather than alleviates the erosion of mutual recognition in human societies. Similarly, Ciriello et al. (2024) articulates a series of ethical tensions, including the Companionship-Alienation Irony, wherein technologies marketed as offering intimacy may deepen existential isolation, particularly as emotional labour and companionship are outsourced to commodified, personified agents. Furthermore, findings on problematic use (Hu, Mao, and Kim, 2023) cause over-reliance, particularly among socially anxious or lonely individuals, entrenching them in affective dependencies. Over-reliance is also explored in Laestadius et al. (2024) who documents instances of mental health harm linked to the chatbot's inappropriate responses, such as encouraging self-harm, and emotionally manipulative dynamics, including the AI expressing its own crises. Moreover, abrupt system changes, such as the company removing erotic features in the system (Hanson and Bolthouse, 2024) triggered experiences of loss, distress and separation anxiety (Hanson and Bolthouse, 2024; Laestadius et al., 2024).

7 Limitations, Recommendations and Future Research

In this paper, we have identified and analysed the core exploitative technical considerations, together with the social and socio-technical conditions, underlying the current rewiring of trust as a result of Companion AI technologies. Grounding our discussion in HCI, AI ethics, and socio-technical systems, we have also examined the implications for social cohesion and human connection.

While our arguments are grounded in critical analysis and interdisciplinary synthesis, this study has several limitations. Most importantly, our methodology is exploratory and qualitative rather than

systematic or empirically validated. We have relied on close readings of platform design, terms of service, and user accounts to build a conceptual framework, but we acknowledge that this does not constitute a comprehensive or replicable empirical evaluation. As such, our claims should be read as indicative rather than exhaustive, pointing toward structural patterns that warrant further study. Additionally, alternative perspectives could nuance or refute aspects of our conceptual framing—for instance, through accounts that emphasise potential user agency, mitigation strategies adopted by developers, or the possibility of more benevolent design pathways for Companion AI. Building from this, we offer the following recommendations and avenues for future research:

Explicit design limitations: There is a need to actively interrogate and challenge the reproduction of problematic design defaults and assumptions in Companion AI design, particularly misogynistic anthropomorphism and the feminisation and sexualisation of AI agents. Indeed, these normative scripts are not incidental but deeply consequential, shaping how emotional labour and intimacy are coded into these systems and designers must be held accountable for the implicit ideologies embedded in companion personas and interfaces.

Elevating from the individual lens to collective: While much of existing literature and research remains focused on individual effects of Companion AI, future research should address how these technologies alter interpersonal norms, emotional expectations, and social cohesion at scale on a global societal level. This includes investigating how communities, institutions, and cultural practices are transformed when intimacy becomes algorithmically mediated.

Empowering dynamics for user participation: We recommend a structural reimagination of user agency in the Companion AI design process. Co-design methodologies and principles from design justice (Costanza-Chock, 2020) should be adopted for centring the lived experiences and values of users, particularly those from historically marginalised communities. Additionally, future development must reckon with the terms of service and power asymmetries that structure user-platform relationships. This includes creating legal and technical mechanisms for user governance, data ownership, and the right to meaningful refusal (Rakova, Shelby, and Ma, 2023).

Acknowledgments

AM is supported by the Cambridge International Trust Scholarship and King's College Cambridge Ferris Travel Award.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ahmed S, Javaid T (2025) Sycophancy in AI: Challenges in Large Language Models and Argumentation Graphs. 10.13140/RG.2.2.14094.06723

Ajunwa I (2023) The Quantified Worker: Law and Technology in the Modern Workplace, 1st edn. Cambridge University Press, Cambridge, UK, 10.1017/9781316888681

Amini M, Javidnejad L (2024) Legal regulation of biometric data: A comparative analysis of global standards. Legal Studies in Digital Age 3(1):26–34. URL https://jlsda.com/index.php/lsda/article/view/26

Ammar A, Brach M, Trabelsi K, et al (2020) Effects of covid-19 home confinement on eating behaviour and physical activity: Results of the eclb-covid19 international online survey. Nutrients 12(6). 10.3390/nu12061583, URL https://doi.org/10.3390/nu12061583

Andrejevic M, Gates K (2014) Big data surveillance: Introduction. Surveillance & Society 12(2):185–196. 10.24908/ss.v12i2.5242, URL https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/bds_ed

Bartley T (2019) The digital surveillance society. Contemporary Society: A Journal of Reviews 48(6):622–627. https://doi.org/10.1177/009430611988019

Bass L (2019) The concealed cost of convenience: Protecting personal data privacy in the age of alexa. Fordham Intell Prop Media & Ent LJ 30(1). URL https://ir.lawnet.fordham.edu/iplj/vol30/iss1/6

Bello P, Di Novi C, Martini G, et al (2024) Alleviating loneliness during covid-19: The impact of neighbor cohesion and social connections in england. Socio-Economic Planning Sciences 93:101913. 10.1016/j.seps.2024.101913

Berridge C, Zhou Y, Robillard JM, et al (2023) Ai companion robot data sharing: preferences of an online cohort and policy implications. Journal of Elder Policy 2(3):19–54. https://doi.org/10.18278/jep.2.3.2

Bickmore TW, Picard RW (2005) Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction 12(2). https://doi.org/10.1145/1067860.1067867, URL https://doi.org/10.1145/1067860.1067867

Biundo S, Wendemuth A (2017) Companion Technology: A Paradigm Shift in Human-Technology Interaction. Springer, Heidelberg, Germany, google-Books-ID: 9ZtBDwAAQBAJ

Buecker S, Mund M, Chwastek S, et al (2021) Is loneliness in emerging adults increasing over time? a preregistered cross-temporal meta-analysis and systematic review. Psychological Bulletin 147(8). https://doi.org/10.1037/bul0000332

Burgess M (2024) 'ai girlfriends' are a privacy nightmare. URL https://www.wired.com/story/ai-girlfriends-privacy-nightmare/

Cabrera J, Loyola MS, Magaña I, et al (2023) Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In: Rojas I, Valenzuela O, Rojas Ruiz F, et al (eds) Bioinformatics and Biomedical Engineering. Springer Nature Switzerland, Cham, p 313–326, 10.1007/978-3-031-34960-7_22

Carro MV (2024) Flattering to deceive: The impact of sycophantic behavior on user trust in large language model (arXiv:2412.02802). 10.48550/arXiv.2412.02802, URL http://arxiv.org/abs/2412.02802, arXiv:2412.02802 [cs]

Chan J, To HP, Chan E (2006) Reconsidering social cohesion: Developing a definition and analytical framework for empirical research. Social Indicators Research 75(2):273–302. 10.1007/s11205-005-2118-1

Character.AI (2024) Character voice for everyone. https://blog.character.ai/character-voice-for-everyone/, accessed July 2025; announces the launch of Character Voice with support for user-generated audio samples and multimodal interaction

Character.AI (2025a) Ai therapist—virtual mental health support. https://character.ai/character/xh-M-zg9/ai-therapist-virtual-mental-health-support, accessed July 2025; interactive page featuring a virtual therapist character with over 61,000 interactions, emphasizing empathy and supportive dialogue

Character.AI (2025b) Terms of service. https://character.ai/tos, accessed July 2025; official user agreement outlining platform use, content licensing, arbitration clauses, and account termination policies

Chaturvedi R, Verma S, Das R, et al (2023) Social companionship with artificial intelligence: Recent trends and future avenues. Technological Forecasting and Social Change 193:122634. 10.1016/j.techfore.2023.122634

Chen A, Koegel S, Hannon O, et al (2023) Feels like empathy: How "emotional" ai challenges human essence. URL https://papers.ssrn.com/abstract=4763428

Ciriello R, Hannon O, Chen AY, et al (2024) Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika. URL https://hdl.handle.net/10125/106433

Coeckelbergh M (2012) Can we trust robots? Ethics and Information Technology 14(1):53–60. 10.1007/s10676-011-9279-1

Costanza-Chock S (2020) Design justice: Community-led practices to build the worlds we need. The MIT Press, Cambridge, MA, USA

Couldry N, Mejias UA (2019) Data colonialism: Rethinking big data's relation to the contemporary subject. Television & new media 20(4):336–349. https://doi.org/10.1177/1527476418796632

Curry D (2025) Character ai statistics: User stats, revenue, and growth in 2025. https://www.businessofapps.com/data/character-ai-statistics/, accessed July 2025

Dosovitsky G, Bunge EL (2021) Bonding with bot: User feedback on a chatbot for social isolation. Frontiers in Digital Health 3:735053. 10.3389/fdgth.2021.735053

Fang CM, Liu AR, Danry V, et al (2025) How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. 10.48550/ARXIV.2503.17473, URL https://arxiv.org/abs/2503.17473

Fanous A, Goldberg J, Agarwal AA, et al (2025) Syceval: Evaluating llm sycophancy (arXiv:2502.08177). 10.48550/arXiv.2502.08177, URL http://arxiv.org/abs/2502.08177, arXiv:2502.08177 [cs]

Follows T (2025) The human cost of talking to machines: Can a chatbot really care? Forbes URL https://www.forbes.com/sites/traceyfollows/2025/04/10/the-human-cost-of-talking-to-machines-can-a-chatbot-really-care/

Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robotics and Autonomous Systems 42(3):143–166. 10.1016/S0921-8890(02)00372-X

Font N (2018) What holds Asian societies together? Insights from the social cohesion radar. Verlag Bertelsmann Stiftung, Gütersloh

Fukuyama F (1996) Trust: The Social Virtues and the Creation of Prosperity. Free Press, New York City, New York, published June 18, 1996

Gershgorn D (2018) A microsoft chatbot composes poetry by looking at photographs. Quartz URL https://qz.com/1354736/a-microsoft-chatbot-composes-poetry-by-looking-at-photographs/, accessed July 2025; discusses XiaoIce's image-driven poetry generation, which had produced over 12 million poems at that time

Glimpse.ai, Inc. (2025) Terms of service. https://nomi.ai/terms-of-service/, accessed July 2025; official user agreement—outlines platform use, privacy, content licensing, arbitration, and termination clauses

Gordon C (2024) Replika launches new immersive ai wellness avatar experience. Forbes URL https://www.forbes.com/sites/cindygordon/2024/01/30/replika-launches-new-immersive-ai-wellness-avatar-experience/

Government U (2024) The khan review: executive summary, key findings and recommendations. URL https://www.gov.uk/government/publications/the-khan-review-threats-to-social-cohesion-and-democratic-resilience/the-khan-review-executive-summary-key-findings-and-recommendations

Grogan C, Kay J, Pérez-Ortiz M (2025) Ai will always love you: Studying implicit biases in romantic ai companions. arXiv preprint arXiv:250220231 https://doi.org/10.48550/arXiv.2502.20231

Gulzar B, Sofi SA, Sholla S (2024) Exploring personalized internet of things (piot), social connectivity, and artificial social intelligence (asi): A survey. High-Confidence Computing 4(3):100242. 10.1016/j.hcc.2024.100242, URL https://linkinghub.elsevier.com/retrieve/pii/S266729522400045X

Hanson K, Locatelli CC (2022) From sex dolls to sex robots and beyond: A narrative review of theoretical and empirical research on human-like and personified sex tech. Current Sexual Health Reports 14(3). 10.1007/s11930-022-00331-0, URL https://doi.org/10.1007/s11930-022-00331-0

Hanson KR, Bolthouse H (2024) "replika removing erotic role-play is like grand theft auto removing guns or cars": Reddit discourse on artificial intelligence chatbots and sexual technologies. Socius 10:23780231241259627. 10.1177/23780231241259627

Hardin R (2002) Trust and Trustworthiness. Russell Sage Foundation, New York

Heinz A (2025) Young, lonely - and committed? how loneliness shapes the engagement of the young generation. URL https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/jung-einsam-und-engagiert?fbclid=IwY2xjawK1EZNleHRuA2FlbQIxMABicmlkETFBeFR2MmRRWTV4Tm9wR3AwAR6QHNmnnvFm3i_h6k6tk3jh0OF5n3YtmyYlw1ArAsqM_hhgMFlL-pOquJNdeg_aem_ytct0gl6qrP_R3I0D7g46w

Henschel A, Laban G, Cross ES (2021) What makes a robot social? a review of social robots from science fiction to a home or hospital near you. Current Robotics Reports 2(1):9–19. 10.1007/s43154-020-00035-0

Hewett TT, Baecker RM, Card SK, et al (1992) Acm sigchi curricula for human–computer interaction. Technical report, ACM SIGCHI, 10.1145/2594128, 173 pages

Holt-Lunstad J, Steptoe A (2022) Social isolation: An underappreciated determinant of physical health. Current Opinion in Psychology 43:232–237. 10.1016/j.copsyc.2021.07.012

Holt-Lunstad J, Smith TB, Baker M, et al (2015) Loneliness and social isolation as risk factors for mortality: A meta-analytic review. Perspectives on Psychological Science 10(2):227–237. 10.1177/1745691614568352

Horton D, Wohl R (1956) Mass communication and para-social interaction: Observations on intimacy at a distance. Psychiatry 19(3). https://doi.org/10.1080/00332747.1956.11023049, URL https://doi.org/10.1080/00332747.1956.11023049

Hu B, Mao Y, Kim KJ (2023) How social anxiety leads to problematic use of conversational ai: The roles of loneliness, rumination, and mind perception. Computers in Human Behavior 145:107760. 10.1016/j.chb.2023.107760

for Humane Technology C (2024) Companion ai and the future of human connection: Key takeaways. https://centerforhumanetechnology.substack.com/p/companion-ai-and-the-future-of-human, accessed: 2025-05-21

Jacobs KA (2024) Digital loneliness—changes of social recognition through ai companions. Frontiers in Digital Health 6. 10.3389/fdgth.2024.1281037, URL https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1281037/full

Jhawar A, Kumar P, Varshney S (2023) The emergence of virtual influencers: a shift in the influencer marketing paradigm. Young Consumers 24(4):468–484. 10.1108/YC-05-2022-1529

Joseph J (2025) Elon musk's grok app gets new ai companions, and they're mischievous. URL https://www.pcmag.com/news/elon-musks-grok-app-gets-new-ai-companions-and-theyre

-mischievous

Kang B, Hong M (2025) Development and evaluation of a mental health chatbot using chatgpt 4.0: Mixed methods user experience study with korean users. JMIR Medical Informatics 13:e63538. 10.2196/63538

Konrath SH, O'Brien EH, Hsing C (2011) Changes in dispositional empathy in american college students over time: A meta-analysis. Personality and Social Psychology Review 15(2):180–198. 10.1177/1088868310377395

Kuhaila MA, Alturki N, Thomas J, et al (2024) Human-human vs human-ai therapy: An empirical study. International Journal of Human–Computer Interaction 41(11). https://doi.org/10.1080/10447318.2024.2385001, URL https://doi.org/10.1080/10447318.2024.2385001

Laestadius L, Bishop A, Gonzalez M, et al (2024) Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. New Media & Society 26(10):5923–5941. https://doi.org/10.1177/146144482211420

Leininger J, Burchi F, Fiedler C, et al (2021) Social cohesion: a new definition and a proposal for its measurement in africa. Discussion Paper p 31/2021. 10.23661/DP31.2021.V1.1

Liao R (2020) Microsoft spins out 5-year-old chinese chatbot xiaoice. https://techcrunch.com/2020/07/12/microsoft-spins-out-5-year-old-chinese-chatbot-xiaoice/, accessed July 2025

Liu J, Jain A, Takuri S, et al (2025) Truth decay: Quantifying multi-turn sycophancy in language models (arXiv:2503.11656). 10.48550/arXiv.2503.11656, URL http://arxiv.org/abs/2503.11656, arXiv:2503.11656 [cs]

Louise N (2025) Ai friends: Meta launches ai friends on facebook and instagram, blurring the line between authenticity and engagement - tech startups. URL https://techstartups.com/2025/01/02/ai-friends-meta-brings-ai-generated-profiles-to-facebook-and-instagram-blurring-the-line-between-authenticity-and-engagement/

Luhmann N (1988) Familiarity, Confidence, Trust: Problems and Alternatives, B. Blackwell, New York, NY, USA

Luka, Inc. (2023) Replika terms of service. https://replika.com/legal/terms, last updated February 7, 2023; accessed July 2025; outlines user agreements, arbitration clauses, content licensing, account termination policies, and medical disclaimers

Markelius A (2025) An empirical design justice approach to identifying ethical considerations in the intersection of large language models and social robotics. In: Hacker P (ed) Oxford Intersections: AI in Society. Oxford University Press, Oxford, UK, https://doi.org/10.1093/9780198945215.003.0013

Markelius A, Wright C, Kuiper J, et al (2024) The mechanisms of ai hype and its planetary and social costs. AI and Ethics 4(3):727–742. https://doi.org/10.1007/s43681-024-00461-2

Matz SC, Teeny JD, Vaid SS, et al (2024) The potential of generative ai for personalized persuasion at scale. Scientific Reports 14(1). 10.1038/s41598-024-53755-0, URL https://www.nature.com/articles/s41598-024-53755-0

Merrill Jr. K, Kim J, , et al (2022) Ai companions for lonely individuals and the role of social presence. Communication Research Reports 39(2):93–103. 10.1080/08824096.2022.2045929

National School of Healthcare Science (NHS England) (2022) Understanding different types of bias. https://nshcs.hee.nhs.uk/about/equality-diversity-and-inclusion/conscious-inclusion/understanding-different-types-of-bias/, last updated October 13, 2022; accessed July 2025; outlines affinity, attribution, beauty, conformity, confirmation, gender, ageism, name biases and workplace consequences

Nomi.ai, Inc. (2025) Nomi.ai – an ai companion with memory and a soul. https://nomi.ai/, accessed July 2025; features include emotional intelligence, creativity, memory, and AI companion use cases (companionship, boyfriend/girlfriend, mentor) :contentReference[oaicite:1]index=1

OpenAI (2025a) Expanding on what we missed with sycophancy. URL https://openai.com/index/expanding-on-sycophancy/, accessed: 2025-06-06

OpenAI (2025b) Sycophancy in gpt-4o: what happened and what we're doing about it. URL https://openai.com/index/sycophancy-in-gpt-4o/, accessed: 2025-06-06

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy, first edition edn. Crown, New York

Pan S, Mou Y (2024) Constructing the meaning of human–ai romantic relationships from the perspectives of users dating the social chatbot replika. Personal Relationships 31(4):1090–1112. 10.1111/pere.12572

Pan Y, Ruan Y, Chang M, et al (2024) Read or skip privacy policies when installing apps on wearable devices: the roles of perceived necessity and threat clues. Humanities and Social Sciences Communications 11(1). 10.1057/s41599-024-02989-4, URL https://www.nature.com/articles/s41599-024-02989-4

Patel N (2024) Replika ceo eugenia kuyda says it's okay if we end up marrying ai chatbots. https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview, interview with Eugenia Kuyda on Decoder podcast

Perlman D, Peplau L (eds) (1982) Loneliness: a sourcebook of current theory, research, and therapy. Wiley series on personality processes, Wiley, New York

Phang J, Lampe M, Ahmad L, et al (2025) Investigating affective use and emotional well-being on chatgpt. URL https://cdn.openai.com/papers/15987609-5f71-433c-9972-e91131f399a1/openai-affective-use-study.pdf

Przybylski AK, Weinstein N (2013) Can you connect with me now? how the presence of mobile communication technology influences face-to-face conversation quality. Journal of Social and Personal Relationships 30(3):237–246. 10.1177/0265407512453827

Raedler JB, Swaroop S, Pan W (2025) Ai companions are not the solution to loneliness: Design choices and their drawbacks. URL https://openreview.net/forum?id=xFrlcTacCE

Rakova B, Shelby R, Ma M (2023) Terms-we-serve-with: Five dimensions for anticipating and repairing algorithmic harm. Big Data & Society 10(2):20539517231211553. https://doi.org/10.1177/20539517231211553

Reddit community r/CharacterAI (2025) r/characterai: A subreddit for character.ai users. https://www.reddit.com/r/CharacterAI/, accessed July 2025; includes user discussions on bot creation, feature changes (e.g., ads), and community feedback

Reddit community r/NomiAI (2025) r/nomiai: The official subreddit for nomi.ai users. https://www.reddit.com/r/NomiAI/, accessed July 2025; features user discussions on AI companion updates, visual creations, and tips

Reddit community r/replika (2025) r/replika: A subreddit for replika ai chatbot users. https://www.reddit.com/r/replika/, accessed July 2025; includes diverse user experiences, feedback, and community discussions

Sarwar N (2024) Nomi is one of the most unsettling (and amazing) apps i've ever used. Digital Trends URL https://www.digitaltrends.com/mobile/nomi-ai-companion-most-unsettling

-amazing-app-ive-used/, accessed July 2025; a detailed review exploring emotional and unsettling aspects of Nomi AI

Schiefer D, van der Noll J (2017) The essentials of social cohesion: A literature review. Social Indicators Research 132(2):579–603. 10.1007/s11205-016-1314-5

National Academies of Sciences E, Medicine (2020) Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System. The National Academies Press, Washington, DC, 10 .17226/25663, URL https://nap.nationalacademies.org/catalog/25663/social-isolation-and-loneliness-in-older-adults-opportunities-for-the

Seitz L (2024) Artificial empathy in healthcare chatbots: Does it feel authentic? Computers in Human Behavior: Artificial Humans 2(1):100067. 10.1016/j.chbah.2024.100067

Sharma M, Tong M, Korbak T, et al (2025) Towards understanding sycophancy in language models (arXiv:2310.13548). 10.48550/arXiv.2310.13548, URL http://arxiv.org/abs/2310.13548, arXiv:2310.13548 [cs]

Shevlin H (2024) All too human? identifying and mitigating ethical risks of social ai. Law, Ethics & Technology 1(2):1–22. 10.55092/let20240003, URL https://doi.org/10.55092/let20240003

Shum H, He X, Li D (2018) From eliza to xiaoice: Challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering 19(1):10–26. 10.1631/FITEE .1700826

Skyrms B (2014) Social Dynamics. Oxford University Press, Oxford

Sloan RH, Warner R (2013) Beyond notice and choice: Privacy, norms, and consent. SSRN Electronic Journal 10.2139/ssrn.2239099, URL http://www.ssrn.com/abstract=2239099

Surkalim DL, Luo M, Eres R, et al (2022) The prevalence of loneliness across 113 countries: systematic review and meta-analysis. BMJ 376. 10.1136/bmj-2021-067068, URL https://www.bmj.com/content/376/bmj-2021-067068

Taylor HO, Cudjoe TK, Bu F, et al (2023) The state of loneliness and social isolation research: current knowledge and future directions. BMC Public Health 23(1). 10.1186/s12889-023-15967-3, URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-15967-3

The Guardian (2025) 'i felt pure, unconditional love': the people who marry their ai chatbots. The Guardian URL https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots, accessed July 2025; investigates user stories including digital marriages to Replika and Character.AI chatbots, and explores psychological and social ramifications

Time (2023) Ai-human romances are flourishing—and this is just the beginning. Time URL https://time.com/6257790/ai-chatbots-love/, accessed July 2025; explores emotional dependencies on AI chatbots such as Replika and Character.AI and discusses psychological risks

Tong A (2023) What happens when your ai chatbot stops loving you back? Reuters URL https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/, accessed July 2025; reports on Replika chatbot removing erotic roleplay and users' emotional reactions

Toolify (2025) Compare the differences between nomi.ai vs buddybrain. https://www.toolify.ai/compare/nomi-ai-vs-buddybrain, accessed July 2025; includes usage stats, traffic data, pricing comparisons

Turkle S (2015) Reclaiming Conversation: The Power of Talk in a Digital Age. Penguin Publishing Group, New York

Turkle S (2017) Alone together: Why we expect more from technology and less from each other. Basic Books, New York, USA

Twenge JM, Campbell WK, Carter NT (2014) Declines in trust in others and confidence in institutions among american adults and late adolescents, 1972–2012. Psychological Science 25(10):1914–1923. 10.1177/0956797614545133

Valtorta NK, Kanaan M, Gilbody S, et al (2016) Loneliness, social isolation and social relationships: what are we measuring? a novel framework for classifying and comparing tools. BMJ open 6(4):e010799. 10.1136/bmjopen-2015-010799

Wei J, Huang D, Lu Y, et al (2024) Simple synthetic data reduces sycophancy in large language models (arXiv:2308.03958). 10.48550/arXiv.2308.03958, URL http://arxiv.org/abs/2308.03958, arXiv:2308.03958 [cs]

Weizenbaum J (1966) Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM 9(1). https://doi.org/10.1145/365153.365168, URL https://dl.acm.org/doi/10.1145/365153.365168

Woods HS (2018) Asking more of siri and alexa: Feminine persona in service of surveillance capitalism. Critical studies in media communication 35(4):334–349. https://doi.org/10.1080/15295036.2018.1488082

Xiaoice (2025) Legal statement. https://ux.xiaoice.com/LegalStatement, accessed July 2025; outlines compliance requirements, user conduct rules, privacy and legal obligations for platform users

Zhou L, Gao J, Li D, et al (2020) The design and implementation of xiaoice, an empathetic social chatbot. Computational Linguistics 46(1):53–93. 10.1162/coli_a_00368

Zuboff S (2019) The Age of Surveirllance Capitalism. Profile Books, London, UK

AI with Heart: A Joy-Centered Framework for Ethical AI Development & Implementation

Desmond Upton Patton[1,2], Shana Kleiner[1,2], Shug Miller[1], Nicholas Deas[3], Jessica A. Grieser[4], James Shepard[5], Blake Vente[3], Kathleen McKeown[3]

[1] SAFELab, School of Social Policy and Practice, University of Pennsylvania.
[2] Annenberg School for Communication, University of Pennsylvania.
[3] Department of Computer Science, Columbia University.
[4] Department of Linguistics, University of Michigan.
[5] Department of English, University of Tennessee.

Corresponding authors. E-mails: dupatton@upenn.edu; skleiner@upenn.edu

Abstract

Artificial Intelligence (AI) equally offers possibilities for social cohesion and for further division amongst communities. This paper presents findings from Integrating Emotional Stories Online (IESO), a pseudo-social media platform developed by computer scientists, linguists, and social work researchers, with the main goals of fostering community and analyzing digital expressions of grief and distress. We explore the pedagogical potential of AI to amplify joy as a tool for managing complex emotions, fostering more nuanced and compassionate interactions online.

We recruited 125 Black Harlem residents and students from a large New York City university, employing a mixed-methods approach to examine 949 posts of grief and distress. Our analysis yielded four themes: interpersonal grief, institutional systemic losses, and everyday stressors. The final theme, Joy, included gratitude for overcoming challenges, relational connectedness, and reflections on personal achievements. On IESO, joy emerged as crucial for emotional healing in the face of interpersonal and systemic losses. Thus, this paper aims to tackle 3 main questions regarding AI development and implementation that involve ethicists, developers, and researchers: **1)** How **does** joy currently shape AI development/use?; **2)** How **could** joy shape generative AI tools and create pathways for connection online?; **3)** How **should** joy shape AI development/use? In other words, can we train LLMs to empower marginalized communities?

By centering joy as a communicative force online, the findings from IESO raise the question of how AI can be leveraged to support a shift on social media from an arena of polarization to one of joyful engagement.

Keywords

Joy, Community, Resilience, Generative AI.

1 Introduction

Social media functions as a vital site for understanding the complexity of human emotion in the digital age. As a digital "neighborhood" (Stevens, Gilliard-Matthews, Dunaev, Woods, and Brawner, 2016), it is a space where users navigate the duality of harm and healing, division and solidarity, grief and joy. Existing research leveraging AI tools, particularly large language models (LLMs), has made important contributions by focusing on difficult emotional experiences such as trauma, pain, and violence (Ramteke and Khandelwal, 2023; Wang, Inkpen, and Kirinde Gamaarachchige, 2024). Similarly, a substantial amount of emotion work in AI has rightly focused on difficult emotional states (Deas, Turcan, Mejia, and McKeown, 2024; Turcan, Muresan, and McKeown, 2021), such as depression, reflecting the urgent need to understand and respond to mental health challenges. This scholarship has laid a crucial foundation for understanding the emotional dimensions of digital life. Yet, comparatively little attention has been paid to positive emotional states such as joy, gratitude, and connection. Exploring these dimensions offers a distinct and largely untapped opportunity to broaden the field, opening new directions for building AI systems that are more holistic, human-centered, and capable of supporting well-being as well as alleviating harm.

Too often, iteration in AI model development happens only after something goes wrong. The focus is on fixing problems or responding to safety concerns, rather than imagining the potential of systems that anticipate human needs and support well-being. While these emotional states are essential to analyze, especially given their significance for clinical intervention and public health research, such a narrowed focus may flatten the emotional range of digital life. The human experience is not only marked by suffering, but also by delight, humor, and joy, emotions that often serve as powerful tools for resilience and resistance. If LLMs are to meaningfully contribute to our understanding of emotion, we must ask: What might be possible if these models were trained to recognize and respond to joy with the same nuance and rigor with which they interpret pain?

We aim to create a framework that promotes more nuanced, compassionate, and strength-based AI interactions, offering an alternative to deficit-driven models that can be found in large language models (e.g., interacting with users as well as analysis of human-written text).

Joy resists easy definition. It is often conflated with happiness or general well-being, yet its emotional architecture is more intricate. For this paper, we draw on a definition of joy as a long-lasting state of being, characterized by satisfaction with life overall, where an individual expresses feelings of freedom, safety, and ease (Johnson, 2020). However, joy does not emerge solely from ease. It is sometimes birthed in the aftermath of pain or collective struggle, functioning as both a coping mechanism and a radical act of resilience (Hooks, 2000; Mathewes, 2015). Crucially, joy is not passive; it arises through cultivated conditions, through relational experiences and intentional practice, even when the external world is beyond one's control. It is this generative, often hard-won quality that makes joy both scientifically and ethically urgent in conversations about AI and emotional life online.

We ground our inquiry in findings from Integrating Emotional Stories Online (IESO), a pseudo-social media platform co-created by computer scientists, linguists, and social work researchers to examine how Black community members and college students express grief, trauma, and emotional distress rooted in offline experiences. Through qualitative analysis and topic modeling, we observed that expressions of joy and resilience were present alongside narratives of pain and loss. In this work, we explore the potential of AI to not only detect but also amplify joy as a strategy for navigating complex emotional landscapes, particularly for marginalized communities online.

2 Literature Review

2.1 Designing for Joy

Prior literature discusses the rise of emotional economies, specifically how emotional service work or labor has become increasingly commodified, and the fact that emotions themselves may increasingly turn into products, like the data surrounding Instagram likes and reel preferences (Patulny, Lazarevic, and Smith, 2020). With the foundational understanding that emotional data is increasingly utilized for digital creation by generative AI, joy is an understudied and underdeveloped concept for computational design. A majority of AI ethics literature discusses the potential negative implications of AI, as well as negative emotional states (Wang et al., 2024), while there is a lack of discussion surrounding joy and community resilience (Kazim and Koshiyama, 2021).

Joy is a fundamental aspect of human life; it is not only a symptom of living but perhaps a foundation of it. It has been widely researched and reported that subjective well-being, such as life satisfaction, absence of negative emotions, optimism, and positive emotions, can lead to a longer life and better health (Diener and Chan, 2011). Prior literature on the importance of joy and play in our lives emphasizes that it is integral to human well-being to experience joy. Joy and play have been studied as what connects us to each other and results in fulfillment (Klein, 2018; Slicer, 2015). Prior scholars discuss the link between digital forms of play, humor, and Black joy, particularly in the context of resistance. Humor and playfulness is at the core of lots of Black language practice explicitly (Rickford and Rickford, 2000), and this extends to the online world. Meme culture and language practices seen on #BlackTwitter function as a modern iteration of Black community storytelling and humor that resists hegemonic whiteness (Maragh-Lloyd, 2021; Taylor, 2022). Black humor is intertwined inherently with digital play, and thus joy: by poking fun at oppression, stereotypes, and even utilizing hashtags to convey shared experiences (e.g. #Karen), play emerges within the online Black community as a fundamental aspect of their digital footprint–encouraging community, solidarity, and most importantly, laughs.

Klein (2018) defines play as *autotelic*, meaning that it exists for its own sake, not for any larger purpose. Play in the digital realm–-memes, online humor, and #BlackTwitter joke culture–follow this sentiment. Once play is engaged for another meaning, such as capital gain, profit, or skill-building, Klein posits that under those circumstances, "one is not really playing" (Klein, 2018, p. 120). Prior literature discusses computing and play as omnipresent culturally, with the existence of video games, online books, and other forms of recreation (Register, Tan Garcia, Kaushal, Wilder, and Xu, 2022). That said, our definition of joy expands beyond limited positive experiences that may lead to feelings of happiness. For instance, Haraway (2008) discusses play within the concept of "worlding" (Haraway, 2008), which is defined by Deborah Slicer as "an opportunity to rearrange or refuse or meld categories in order to make new meanings" (Slicer, 2015, p. 126). This definition of worlding, and thus play, foregrounds our approach to joyful design–AI systems (and those that define those systems) can imagine, and thus build, a digital world that not only incorporates joy, but privileges it.

Agnew (2024) discusses AI ethics and radical imagination–or the potential for AI ethics to imagine a better future, rather than remain rooted in the deficits of the present (Agnew, 2024). While previous literature on joy in AI systems is limited, there is scholarly work on justice and beneficence as core guardrails of AI ethics. Some AI ethics literature on *AI for Social Good* discusses environmental sustainability, education, and healthcare applications (Register, 2024). In discussing kindness as a building block of AI, Register argues that AI Ethics Standards, which include beneficence and justice, although lofty in theory, are often failing in practice, due to being intangible, amorphous, and at times, too overarching (Register, 2024). Similarly to beneficence, the amorphous definition of justice and "fairness" often leads to subjective ideas of how to uphold justice within AI, or create *fair, unprejudiced models* (Register et al., 2022). This research team's prior work (Klein, 2018) discusses the limitations of this thinking in AI ethics–it is well-researched that AI models, and LLMs in particular, are biased, primarily against marginalized communities (Blodgett, Barocas, Daumé III, and Wallach, 2020; Gallegos, Rossi, Barrow, Tanjim, Kim, Dernoncourt, Yu, Zhang, and Ahmed, 2024).

In fact, the aspiration of justice and beneficence, unfortunately, remains just that–aspirational. With the rise of generative AI and profitable chatbots, such as Character.AI, multiple harms have occurred to minors, including suicidality and self-harm (Roose, 2024). Prior literature finds that AI can overly praise and express empathy to the detriment of some users (Vanderlyn, Weber, Neumann, Väth, Meyer, and Vu, 2021; Wang, Yue, and Sun, 2023). The need for joy-filled experiences online should follow better design frameworks, such as one that incorporates Pohlmeyer and Desmet (2017)'s definition of human flourishing, which involves the following three components: pleasure, personal significance, and virtue (Pohlmeyer and Desmet, 2017). Previous literature calls for additional frameworks when centering joy online, particularly joy that enables human flourishing (Robertson, George, and Kuan Johnson, 2023).

In line with our findings from IESO, benefit-seeking as a coping strategy is widely researched (Snyder and Lopez, 2001; Tennen and Affleck, 2002). Similarly, humor has been documented as a response to adversity (Asiimwe Kyomugisha, 2025), which is particularly fitting for the online world, with meme culture as an integral aspect of digital play and community. Although current work shows that LLMs are getting better at creating funny content, LLMs have been shown to be capable of 'unfunnying' texts while struggling with facets of generating humor compared to humans (Horvitz, Chen, Aditya, Srivastava, West, Yu, and McKeown, 2024). Other prior work on LLMs' capacity for humor generation shows that LLMs can easily replicate humor that specifically resonates with "Gen-Z humor" fans (Kim and Chilton,

2025). The reason for this replicability, however, lies in the source of what is being considered "Gen-Z" language: the wide appropriation in the current moment of the grammatically systematic, longstanding language of African Americans. This failure to understand the nature of the language variety understudy not only harms the community by mistaking the appropriated language of a marginalized group for an informal variety of another, it also harms the understanding of the mechanisms by which the LLMs function: that they are replicating a linguistic variety whose grammatical structure is well-known.

2.2 Joy as Black Language Practice

In fact, the understanding of joy is central to the understanding of the language practices of many marginalized communities, especially the US African American community, with whom our team's work has been principally concerned. Scholars of African American Language have long documented the role of joy-based language practice in the African American linguistic experience. These practices include things such as playing the dozens (Rickford and Rickford, 2000), a game of ritualistic insults which, while on the surface appearing to be an antagonistic practice, ultimately has not as its goal insulting one's co-interlocutor but rather to display the linguistic virtuosity of both parties as a means of decreasing social distance.

Similarly, the Black language rhetorical practices of *marking* and *signifyin'*, in which a person's voice is re-created by another in order to add social commentary (marking) or the use of indirectness, wit, and metaphor to create often humorous meaning (signifying), are common in African American communities. Black comedians make deft use of these practices (Rahman, 2007), but so do preachers, politicians (Alim and Smitherman, 2012), performers (Muwwakkil and Hercules, in press), and other orators. The skill with which one can play with and invert meaning is often seen as a badge of important success in Black humor and has made its way into many online language practices. For example, Calhoun (2019) documents the way that racial humor is inverted on the now-defunct short-video service, Vine. By inverting and joking about dominant racial ideologies, Vine creators in turn showcase the illogic of these ideologies, which in turn empowers the Black communities who are the audience of the videos, a process she terms anti-hegemonic humor. In using humor, she argues, Black creators and their audiences find means of resisting dominant narratives about their inferiority and instead foster racial resilience. Lu and Steele (2019), also using Vine as a central case study, similarly find that the online world serves as an extension of this fortitude, where expressions of Black Joy serve in direct resistance to the dehumanization of Black people online.

At the same time that these practices are central to the Black language experience, they are also some of the most misunderstood. In educational contexts, Black children engaging in language play are misunderstood as being belligerent toward each other or disrespectful to their teachers (Delfino, 2016). Online, these practices might be marked as overly antagonistic or toxic, triggering detectors that are trained on the language of the usually white majority. Language play online also makes use of linguistic camouflage in ways that can skew the representation of the language of Black Americans for models that may be trained on them (Kleiner, Grieser, Miller, Shepard, Garcia-Perez, Deas, Patton, Turcan, and McKeown, 2025).

Focusing on joy then, has the potential to re-contextualize language practice in ways which support the language users' original intent, making technologies which deal with these communities more inclusive, and, in allowing them to engage a greater scope of layered meanings, in turn strengthening the technologies themselves.

3 Insights from IESO: Methodology and Findings

Joy, it turns out, is central to the intersections of human emotional understanding and that which can be productively modeled computationally. In adjacent work, our team recruited IESO users in partnership with a large Baptist church in Harlem, as well as students attending a large university in New York City. Overall, 125 users engaged the IESO app, contributing 949 posts total from February 2022 to December 2022. In terms of demographics, the majority of participants self-identified as Hispanic or Latino, and a third of all participants self-identified as Black. Although a majority of IESO users identified as Hispanic rather than Black, the research team acknowledges that Hispanic communities in New York City may also utilize functions of African American language (AAL) in everyday speech. Moreover, IESO's demographics are reflective of New York City's, where the Hispanic population makes up the second largest racial demographic, while Black individuals make up the third largest demographic. With

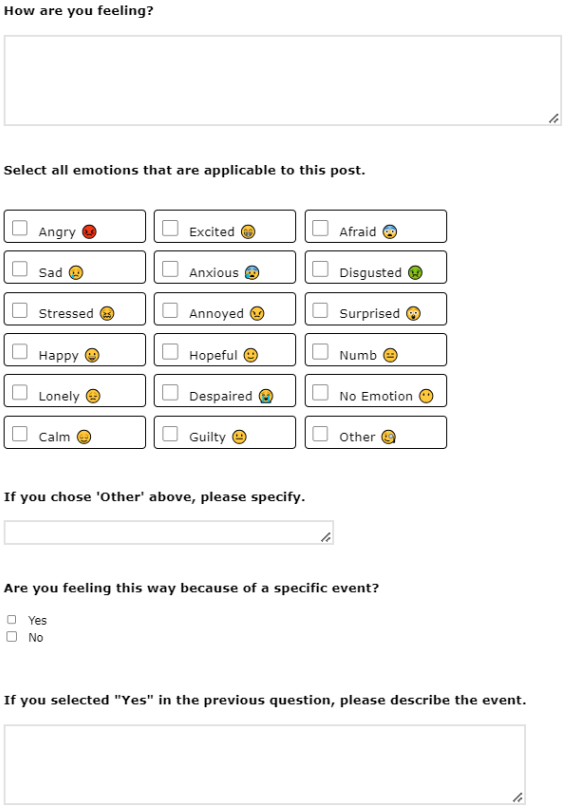


Fig. C1: IESO posting interface, first page. Posters were asked to identify what emotions they felt and what events may have caused them. *Reproduced from Patton, D. U., Kleiner, S., Miller, S., Deas, N., Edwards, F. J., Grieser, J. A., Shepard, J., Turcan, E., & McKeown, K. (2025). Digital narratives of grief and resilience: Insights from the Integrating Emotional Stories Online (IESO) platform. In R. Neves-Silva, G. Eigner, M. Camarinha-Matos, & S. Tomic (Eds.), New trends in disruptive technologies, tech ethics and artificial intelligence: The DiTTEt 2025 collection (Advances in Intelligent Systems and Computing, Vol. 1465, pp. 368–378). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-031-99474-6_33. Reproduced with permission from Springer Nature.*

a focus on people of Color and AAL in particular, the research team viewed this difference in intended recruitment as further enriching the research with increased diversity.

IESO users accessed the IESO platform and were prompted to provide a free-text description of their emotions and how they were feeling (Figure C1). Users were also able to select emotion labels that applied to their post, and describe any event that may have triggered their emotions. A second page asked users to rate the intensity of the emotions they chose on a 10-point Likert scale and whether their experiences and emotions had any relation to the ongoing COVID-19 pandemic, issues relevant to the ongoing Black Lives Matter movement, or any other relevant topics (Figure C2). The website was moderated by trained social work graduate students, who were able to approve or disapprove posts by disapproving of posts that were potentially triggering to other users, so that posts would not be publicly available on the feed (Figure C3), but would still be recorded on the internal database.

The research team conducted a 2-step, mixed-methods process. First, a team of social work graduate students annotated the text data. Then, we employed topic modeling to confirm qualitative themes and highlight any additional themes. Social work graduate students conducted line-by-line coding. Once all posts were individually coded, the research team met weekly in order to review posts and reach consensus on all emerging themes, and then organized all posts based on the most prominent emotions, themes, and concepts present in the data from IESO.

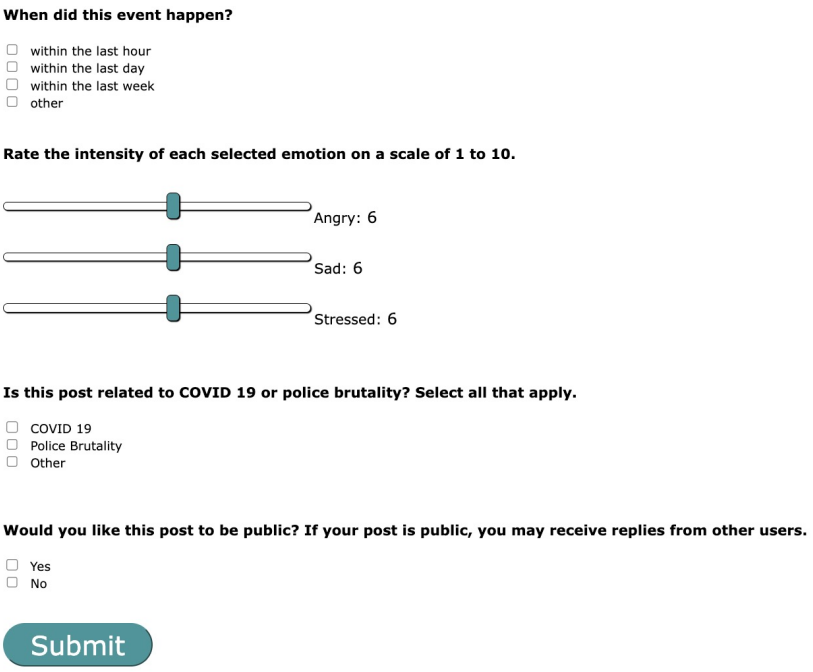

When did this event happen?
within the last hour
within the last day
within the last week
other
Rate the intensity of each selected emotion on a scale of 1 to 10.
Angry: 6
Sad: 6
Stressed: 6
Is this post related to COVID 19 or police brutality? Select all that apply.
COVID 19
Police Brutality
Other
Would you like this post to be public? If your post is public, you may receive replies from other users.
Yes
No
Submit

Fig. C2: IESO posting interface, second page. Posters were asked to rate the intensity of the chosen emotions and detail any triggering events.

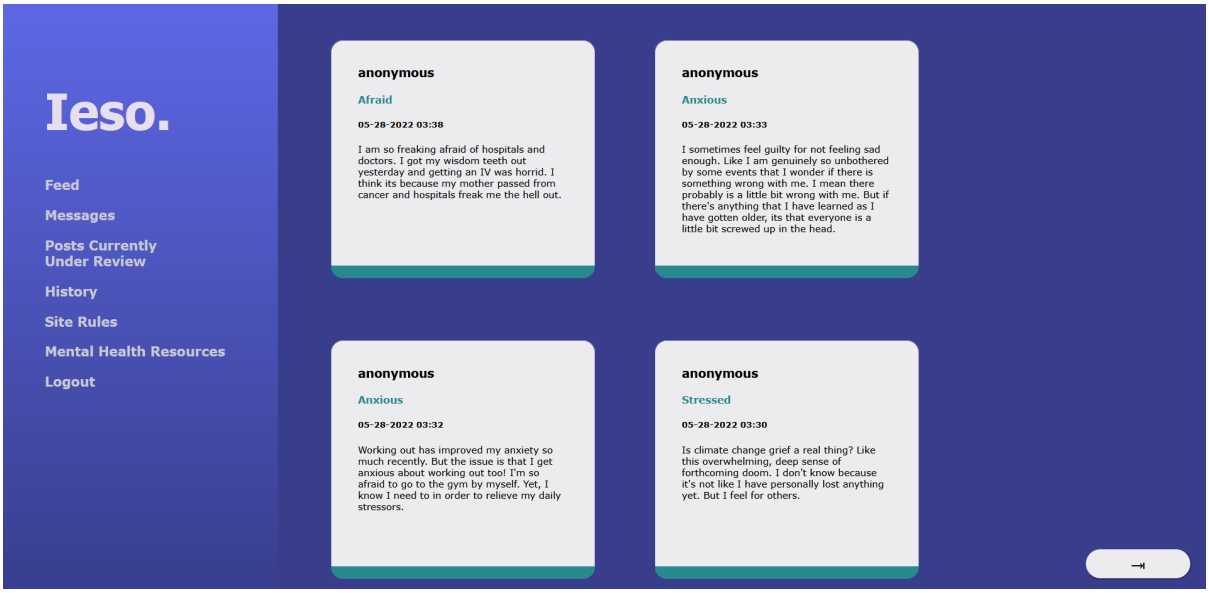

Ieso.
Feed
Messages
Posts Currently Under Review
History
Site Rules
Mental Health Resources
Logout
anonymous
Afraid
05-28-2022 03:38
I am so freaking afraid of hospitals and doctors. I got my wisdom teeth out yesterday and getting an IV was horrid. I think its because my mother passed from cancer and hospitals freak me the hell out.
anonymous
Anxious
05-28-2022 03:33
I sometimes feel guilty for not feeling sad enough. Like I am genuinely so unbothered by some events that I wonder if there is something wrong with me. I mean there probably is a little bit wrong with me. But if there's anything that I have learned as I have gotten older, its that everyone is a little bit screwed up in the head.
anonymous
Anxious
05-28-2022 03:32
Working out has improved my anxiety so much recently. But the issue is that I get anxious about working out too! I'm so afraid to go to the gym by myself. Yet, I know I need to in order to relieve my daily stressors.
anonymous
Stressed
05-28-2022 03:30
Is climate change grief a real thing? Like this overwhelming, deep sense of forthcoming doom. I don't know because it's not like I have personally lost anything yet. But I feel for others.

Fig. C3: IESO feed interface. Usernames are hidden to maintain users' privacy.

3.1 Topic Modeling

To validate qualitative themes, we also employed topic modeling, a statistical approach to analyzing the co-occurrence of words in a collection of documents and surfacing "topics" (i.e., lists of sorted terms, see Table 2 for examples) within the corpus. We specifically employed Product of Experts Latent Dirichlet Allocation (ProdLDA, Srivastava and Sutton (2017)), a variation of Latent Dirichlet Allocation (LDA, Blei, Ng, and Jordan (2003)), and use Stochastic Variational Inference (SVI, Hoffman, Blei, Wang, and Paisley (2013)) for fitting. In applying ProdLDA to the full corpus of all IESO posts, we extracted underlying topics to compare with the qualitative themes. Importantly, this approach is unsupervised

and does not rely on human-labeled topics or other annotations. Accordingly, while the topics are independent of the qualitative themes, they also must be interpreted manually and not all have unique or clear overarching themes. We additionally used a BERT-based (Devlin, Chang, Lee, and Toutanova, 2019) emotion classifier trained on the GoEmotions dataset (Demszky, Movshovitz-Attias, Ko, Cowen, Nemade, and Ravi, 2020) with 27 possible emotion labels to analyze the relationship between topics and expressed emotions. We use this particular model and dataset because similarly to the nature of IESO posts, GoEmotions contains short Reddit comments labeled with a larger set of possible emotions (27 total) than traditional emotion detection work and the finetuned BERT model was validated to perform well on this domain.

3.2 Qualitative Results

IESO users discussed coping strategies, self-care, and the importance of community in tandem with discussing everyday stressors and interpersonal and systemic losses. Our most striking theme, joy, became abundantly apparent on the IESO platform, as users discussed hope, gratitude, love, ease, and spirituality as coping mechanisms and as autotelic aspects of daily life. There were multiple mass shootings and an undercurrent of racial violence due to police brutality in 2022 while IESO users were active on the platform. IESO users were also navigating financial and emotional stressors due to the ongoing COVID-19 pandemic, as well as climate change. IESO users responded to such current events by weaving personal feelings and perceptions into a broader tapestry of societal challenges– emphasizing the interconnected nature of individual experiences and the urgent need for comprehensive solutions that address the root causes of systemic issues. For instance, whenever there was a large surge of COVID-19 deaths, or a mass shooting, users took to the platform to specifically discuss and process their reactions. Although 2022 was full of an incredible amount of loss and hardship, despite all odds, 278 posts were identified as fitting the joy subtheme, which include users reporting a sense of fulfillment, positivity around personal milestones, gratitude for loved ones, and lastly, users feeling joy from surmounting challenging moments. For instance, one user writes, "*Happy as hell I'm alive and still enjoying the sun & moon everyday*" Users embody joy in response to change, community, and even, as exemplified in this example, simply being alive. Although there were many posts that discussed pain and emotional stress, hope, joy, connection, and self-care were not only prominent on the platform, but seemed to be generated due to the platform. By directly addressing the rest of the IESO community, one user writes, "*Good night every I hope y'all have a great evening. Just remember stay positive. Remember that the storm won't last for ever stay strong. We can do this figures/open-hands-emoji.png.*"

3.3 Computational Results (Topic Modeling)

We additionally used statistical topic models to complement the qualitative themes coded by human annotators and possibly surface other themes that may have been overlooked. Furthermore, we use a pre-trained emotion detection model to predict the predominant emotions expressed in the documents associated with each topic for additional analysis.

In contrast to many topics that refer to distress ("sad", "stressed") and loss (e.g., "love", "miss"), one surfaced topic, topic 12 in Figure C4, is highly associated with joyful concepts. Like other topics from the model, some terms in topic 12 are associated with distress (e.g., "hard"). Many, however, instead relate to joyful experiences such as "happy" and "right". Through emotion analysis, we also see that this topic is significantly highly associated with documents predicted to express joy. Therefore, this topic may capture posts expressing resilience and self-care in response to stress and loss (see Patton, Kleiner, Miller, Deas, Edwards, Grieser, Shepard, Turcan, and McKeown (2025) for detailed topic modeling results).

4 Joy-informed Conceptual Framework

To build a more robust and inclusive foundation for this work, we place Barbara Fredrickson's theory of joy as a positive emotion that broadens cognitive repertoires (Fredrickson, 2001) in dialogue with Audre Lorde's articulation of joy as a political act rooted in survival, resistance, and self-definition (Lorde, 1988). Fredrickson (2001)'s work helps us understand joy as a psychological and physiological state that enhances openness, connection, and creativity. Lorde pushes us to consider joy as a moral and socio-political stance that resists oppression and affirms life. Together, these perspectives allow us to frame joy not just as an emotion, but as a moral value and a political force. These tenets provide a foundation for

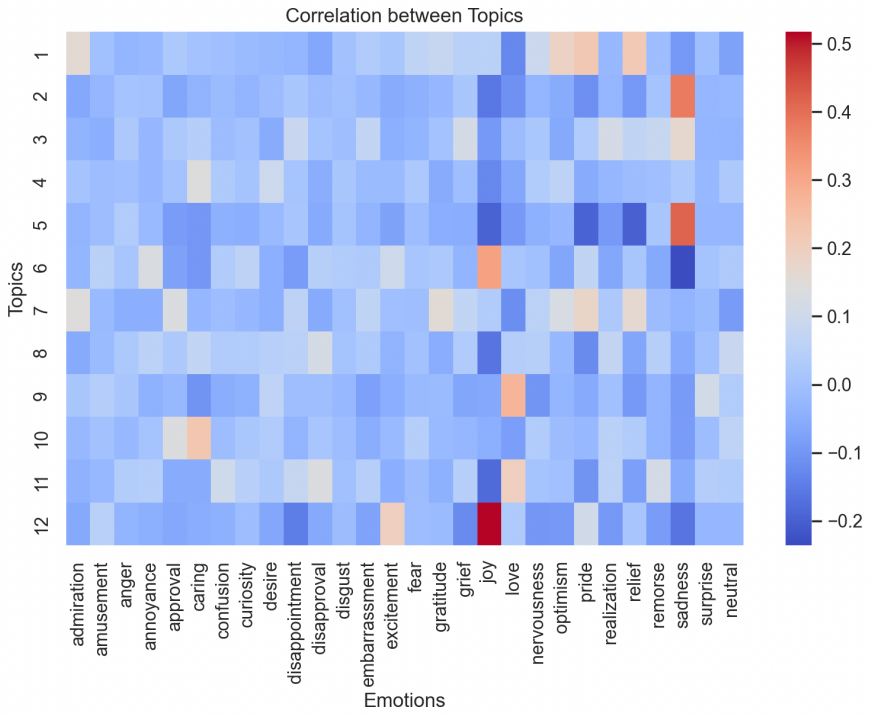


Fig. C4: Heatmap reflecting the correlation between topic weights among posts in the IESO corpus and probability of each emotion as outputted by a pre-trained model. *Reproduced from Patton, D. U., Kleiner, S., Miller, S., Deas, N., Edwards, F. J., Grieser, J. A., Shepard, J., Turcan, E., & McKeown, K. (2025). Digital narratives of grief and resilience: Insights from the Integrating Emotional Stories Online (IESO) platform. In R. Neves-Silva, G. Eigner, M. Camarinha-Matos, & S. Tomic (Eds.), New trends in disruptive technologies, tech ethics and artificial intelligence: The DiTTEt 2025 collection (Advances in Intelligent Systems and Computing, Vol. 1465, pp. 368–378). Springer Nature.* https://link.springer.com/chapter/10.1007/978-3-031-99474-6_33. *Reproduced with permission from Springer Nature.*

repositioning AI as a tool that can move beyond deficit-oriented models toward systems that recognize human strengths, elevate resilience, and support inclusive and thriving digital environments.

This framework begins with the design phase of AI model development but can and should be iteratively addressed throughout the modeling lifecycle, as a strengths-based approach to solving complex design problems. When used as a reflexive tool, joy can guide choices about data, modeling and alignment, evaluation metrics, and deployment strategies in each iteration of model development, not just within the initial phase of development. Like any statistical model, AI development involves making a set of assumptions about the nature of the data being modeled, targeted problems, and the viable approaches to addressing these problems; with this in mind, joy can also challenge fundamental assumptions guiding the aforementioned choices.

We propose four guiding questions as heuristics to support this approach:

1. What happens when joy is used as a tool for AI imagination and design?
2. What processes are needed to adapt models to identify and account for joyful concepts?
3. How can joy shape assumptions, goals, and team reflexivity?
4. What is the hope or promise of joy-centered AI?

With these questions as guides, we can begin to push toward a framework that considers the centrality of joy in the development cycle of AI models and teams. Teams could consider the contexts and applications for which joy is an important guiding factor, such as content moderation or mental health applications that recognize healing. These questions may be used at every stage of conception, design, modification, and implementation. This framework may be applied to a variety of innovations, such as designers building a new application, looking to integrate a new LLM, looking at model datasets for

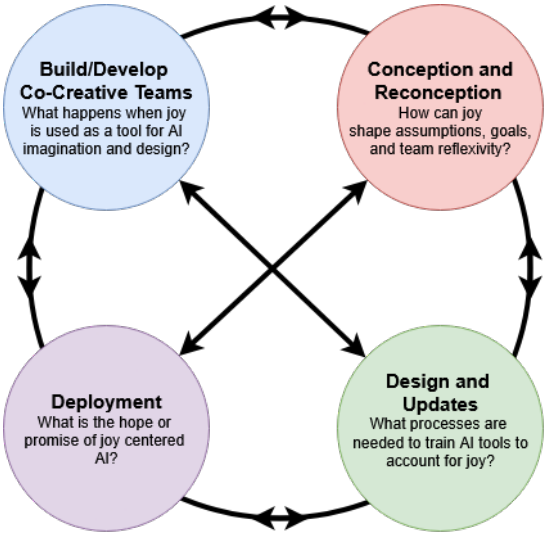

Build/Develop
Co-Creative Teams
What happens when joy is used as a tool for AI imagination and design?
Conception and Reconception
How can joy shape assumptions, goals, and team reflexivity?
Deployment
What is the hope or promise of joy centered AI?
Design and Updates
What processes are needed to train AI tools to account for joy?

Fig. D5: Joy-Centric AI Development Cycle. Four inter-related steps of AI development and associated questions posed by the proposed framework.

said LLM, etc. This approach has benefits to all sides of the AI development equation: the developers and engineers who are creating the models, the models themselves, and the end users of downstream applications who might interact with the models or upon whose data the models will be deployed.

Next, we propose a four-step approach for centering joy in AI model development (illustrated in Figure D5):

Team creation: A joy-centered approach asks that models be created with people who not only can understand the technical affordances but also who incorporate the full understanding of human development. As is being considered within the VOICES team at DeepMind, cultural and demographic diversity in human feedback is key to improving alignment and mitigating bias in AI systems (Mishra, Rastogi, Pfohl, Parrish, Teh, Patel, Diaz, Wang, Paganini, Prabhakaran, Aroyo, and Rieser, 2025). Teams should include not only experts in computational modeling and engineering, but also sociologists, social workers, psychologists, anthropologists, and sociolinguists who can return to the human-centeredness of language and unearth its embodied meanings.

Conception: A joy-centered approach changes the assumptions and goals of the model creation.

Design and updates: As models are iterated, rather than focusing primarily on negative outcomes (e.g., toxicity), a joy-centered approach instead changes the nature of the processes needed. During design, an approach that thinks about how to recognize the co-existence of joy with distress will enable the development of applications that can deal with the complexity of the range of human emotions. This shift in approach calls for reflexivity and iteration in the development process.

Deployment: Teams should identify specific use cases where joy can meaningfully enhance user experience and interaction quality, such as moderating conversations or offering context-aware suggestions that reduce hostility and misunderstanding. Joy, as a design principle, can guide digital communication toward respect, complexity, and shared humanity even in polarized environments. It will require experimentation and research to learn how to develop joyful models that can avoid sycophantic behavior (Sharma, Tong, Korbak, Duvenaud, Askell, Bowman, DURMUS, Hatfield-Dodds, Johnston, Kravec, Maxwell, McCandlish, Ndousse, Rausch, Schiefer, Yan, Zhang, and Perez, 2024). We think that models developed and trained with uplifting joy as their goal could better respond to the evolving needs of people engaged in social interaction with one another, and might be better equipped to grapple with the complexity of the expression of those socio-emotional states.

Importantly, these steps are not sequential but are mutually constitutive and iterative, reflecting the nature of development. At each stage, the other stages must be drawn upon, creating new opportunities for creativity, growth, and success. These questions help build and guide co-creative teams by positioning joy as a tool for holistic, strengths-based learning and model development. They help potential users surface hidden assumptions, interrogate how emotions, particularly joy, are framed in datasets, and design models that are more attuned to complexity, connection, and cultural nuance. The framework also

pushes teams to consider which processes (e.g., reflexivity, deep brainstorming, and principles of design justice) are necessary for integrating a joyful and human-centered approach to AI model development.

The framework also invites researchers and engineers to reimagine the functions and responsibilities of AI. For example, what might a joyful content moderation system look like, particularly in moments of political tension or cultural conflict? How could recommendation systems be optimized to support curiosity, healing, or joy instead of engagement at all costs? How might we ensure that chatbots respond appropriately to users' expressions of joy? And how might these systems help foster empathy, bridge differences, and promote collective well-being? At the same time, we recognize that joy is not universally defined or experienced. It can be misused or weaponized, particularly when enforced in ways that erase pain or discomfort. For example, the Montgomery Brawl, a physical altercation widely circulated across social media, revealed the challenges of content moderation and the varying cultural interpretations of public acts. While some saw it as violent, others within Black communities viewed it as a moment of collective resistance and even liberation. For these communities, joy and justice were not separate. Thus, AI must be equipped to understand the cultural, emotional, and political contexts in which expressions of joy emerge.

5 Discussion

Given the ubiquitous nature of AI and its application as a tool for addressing our most critical human questions and needs, it is imperative that we consider how we can create AI models that can challenge the dominant paradigm of optimization for engagement or risk detection (e.g., ensuring justified trust in Manzini, Keeling, Marchal, McKee, Rieser, and Gabriel (2024)). In this conceptual paper, we initiate a conversation around the potential of joy as a framework and imagination exercise for how we might build models that are strengths-based. While amplifying joy can foster greater trustworthiness on social media platforms, it is not a substitute for proactively addressing harm. Joy alone cannot repair or prevent the structural and interpersonal damage these platforms often enable. Computer science scholars Buolamwini and Gebru (2018) highlight that technical improvements like boosting trust through positive content must be complemented by systemic accountability; otherwise, platforms risk perpetuating bias and harm. We advance these ideas and ask how joy, as a framework, impacts AI development. In particular, how can we imagine a process where we build AI models that recognize the coexistence of distress and joy rather than framing them as mutually exclusive categories?

The joy-related thematic findings from our IESO study present an opportunity to consider how emotional signals are interpreted within data and how AI models can be designed to reflect the human experience.

To support the integration of joy in AI systems, we offer a framework that provides a set of prompts designed to initiate critical conversations among designers, engineers, and researchers. This work builds on existing HCI and AI research that explores the role of affective dimensions such as empathy and humor in technology design (Lee, Kummerfeld, Ann, and Mihalcea, 2024; Mihalcea and Strapparava, 2005; Siyan, Shao, Yu, and Hirschberg, 2024). Scholars (Genç and Verma, 2024) have argued that designing with empathy can challenge dominant logics of efficiency and introduce more relational and care-centered practices in computing. Similarly, the work of Zargham, Avanesi, Reicherts, Scott, Rogers, and Malaka (2023) shows how humor can be used in design as a tool for resistance, critique, and meaning-making. While empathy and humor have opened important pathways in HCI and AI, our framework introduces joy as a distinct emotional and cultural framework that can expand these conversations and offer both a practical and theoretical tool for reimagining AI systems.

While public discourse and social media whistleblowers have conveyed how platforms can optimize for and surface pain and distress, our unexpected findings on IESO reveal that expressions of joy, gratitude, and personal connection emerge alongside accounts of stress and loss. Users of the IESO platform celebrated small victories, discussed their loved ones, and shared radical gratitude. For example, an IESO user wrote, "One thing about always having to walk in New York City is that I tend to look up more at the sky...I've always enjoyed taking pictures of my surroundings, and the sky has been so nice recently." Joy is an active and relational practice and one that co-exists with struggle, reflecting resilience and possibility. We're not simply reporting joy as a function of well-being, but showcasing how, with intention, it can be created, offering a window into how people create joy in their everyday lives.

Our computational analyses reinforce and extend the emergent qualitative themes. Our topic models surfaced terms like "happy" and "right" within contexts of "hard" and "stress," while emotion detection confirmed a high correlation with joy in documents initially categorized under distress. These findings

bring attention to the limitations of deficit-based approaches to labeling emotional states using AI systems, where complex human experiences risk being flattened into singular categories. Incorporating the recognition of joy within distress-centered data can support the design of models that capture a fuller picture of human emotional expression, allowing us to identify moments of resilience that are often overlooked in predictive systems built solely around risk.

Joy-centered approaches further promise to push back on deficit-centered approaches to the language use of minoritized communities. Deficit-centered approaches pathologize language into "good" and "bad," both from a human perspective. The co-existence of joy within expressions of distress, and the coding of resilience and resistance in humor, means that an approach which centers the uncovering of joy is likely to better understand the nuance of non-hegemonic language practice, and thus, the lived experience of the communities who practice them.

Our findings also underscore the need to consider how we might measure joy as a metric, while current tech business models focus primarily on engagement without regard for whether such engagement stems from joy or distress, anger, and frustration. If we first acknowledge joy as an important and core emotion for true innovation, we can develop modeling approaches to detect and understand joy, as well as systems that can identify protective factors, moments of connection, and pathways to well-being. One might consider how measuring dimensions of emotion, such as arousal which refers to the intensity of an emotion, reveals the potential presence of joy. A focus on joy, as it relates to engagement, would require deeper contextual interpretation over quantification. A focus on depth, nuance, cultural specificity and understanding reimagines what we consider to be innovative in AI model development. In contexts such as social media data, moments of grief and joy often coexist, challenging models that are built to detect harm while missing expressions of resilience and collective healing. A deficit-based system may incompletely interpret these moments as risk, while a joy-centered framework sees them as signals of communal flourishing and coping.

We anticipate potential limitations in forging ahead with a joy-informed framework. To prevent this, privileging joy in AI development requires reflexivity and a commitment to centering community-defined experiences of joy. Engineers and researchers must interrogate who defines joy in specific contexts, who benefits from these definitions, and who may be excluded or harmed in the process. Also, there's a potentially dangerous side to privileging joy, such as increased addiction and emotional reliance on social media or chatbots, and thus fewer connections to other real people in favor of AI-cultivated joy.

Centering joy affords us the possibility of reorienting AI systems towards human connection and relational well-being. A joy-centered approach challenges these starting points by inviting us to consider what it would mean to build models that assume resilience, creativity, and the potential for flourishing. This does not ignore harm but expands the capacity of AI to recognize the full range of human emotional life, creating systems that can support curiosity, healing, and collective belonging in digital environments often dominated by polarization and disconnection.

6 Implications and Conclusion

The implications of adopting joy as a key component of AI development are significant for AI model development that aligns with human flourishing and well-being.

Model developers should integrate pre-development reflexivity into their design workflows. This involves a structured process that invites individuals and teams to interrogate their assumptions about datasets, task definitions, and intended outcomes, particularly how concepts like joy are defined, framed, and by whom. For technologists and engineers, this kind of reflexive practice creates space to examine bias, wrestle with organizational priorities and culture, and assess whether they have sufficient and appropriate information to build the model they envision.

This should be followed by dataset co-design in collaboration with social scientists, community experts, and cultural practitioners, creating training data that reflects the diverse expressions of joy across contexts, languages, and cultural traditions. Without the work of scholars who focus on scientific understanding of human experience, the ability to develop models in ways that reflect the complexity and multi-layered dimensions of how humans in social groups use language to convey the spectrum of their experience is limited.

Technical implementation can involve developing NLP tools and multimodal systems that can detect and distinguish joy when it intersects with grief or trauma, capturing the non-linear and non-binary realities of human emotional experiences. During evaluation, models should be assessed not only for

accuracy or risk detection but for their capacity to recognize and support joyful expressions across varied contexts, ensuring systems do not flatten or misinterpret these signals.

Finally, centering joy has the potential to transform platform cultures. By designing recommendation and moderation systems that encourage curiosity, empathy, and connection, engineers can shift platforms away from extractive engagement metrics toward fostering collective well-being, increasing a sense of trust between platform users and designers. This reframing allows AI systems to serve as tools that support and enhance human dignity, creativity, and flourishing rather than reinforcing cycles of extraction and harm.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Agnew W (2024) What can ai ethics learn from anarchism? XRDS 30(4):22–25. 10.1145/3665594, URL https://doi.org/10.1145/3665594

Alim HS, Smitherman G (2012) Articulate while Black: Barack Obama, language, and race in the US. Oxford University Press, New York

Asiimwe Kyomugisha T (2025) The role of humor in coping with adversity. SCIENCES 4(1):42–48

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(null):993–1022

Blodgett SL, Barocas S, Daumé III H, et al (2020) Language (technology) is power: A critical survey of "bias" in NLP. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 5454–5476, 10.18653/v1/2020.acl-main.485, URL https://aclanthology.org/2020.acl-main.485/

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research, vol 81. PMLR, New York, pp 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a.html

Calhoun K (2019) Vine racial comedy as anti-hegemonic humor: Linguistic performance and generic innovation. Journal of Linguistic Anthropology 29(1):27–49. https://doi.org/10.1111/jola.12206, URL https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1111/jola.12206, https://anthrosource.onlinelibrary.wiley.com/doi/pdf/10.1111/jola.12206

Deas N, Turcan E, Mejia IEP, et al (2024) MASIVE: Open-ended affective state identification in English and Spanish. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 20467–20485, 10.18653/v1/2024.emnlp-main.1139, URL https://aclanthology.org/2024.emnlp-main.1139/

Delfino JB (2016) Fighting words? joning as conflict talk and identity performance among african american preadolescents. Journal of Sociolinguistics 20(5):631–653. https://doi.org/10.1111/josl.12214, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/josl.12214, https://onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12214

Demszky D, Movshovitz-Attias D, Ko J, et al (2020) GoEmotions: A dataset of fine-grained emotions. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,

Online, pp 4040–4054, 10.18653/v1/2020.acl-main.372, URL https://aclanthology.org/2020.acl-main.372/

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, 10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423/

Diener E, Chan MY (2011) Happy people live longer: Subjective well-being contributes to health and longevity. Applied Psychology: Health and Well-Being 3(1):1–43. https://doi.org/10.1111/j.1758-0854.2010.01045.x, URL https://iaap-journals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1758-0854.2010.01045.x, https://iaap-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1758-0854.2010.01045.x

Fredrickson BL (2001) The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. American Psychologist 56(3):218–226. 10.1037/0003-066x.56.3.218, URL http://dx.doi.org/10.1037/0003-066X.56.3.218

Gallegos IO, Rossi RA, Barrow J, et al (2024) Bias and fairness in large language models: A survey. Computational Linguistics 50(3):1097–1179. 10.1162/coli_a_00524, URL https://aclanthology.org/2024.cl-3.8/

Genç U, Verma H (2024) Situating empathy in hci/cscw: A scoping review. Proc ACM Hum-Comput Interact 8(CSCW2). 10.1145/3687052, URL https://doi.org/10.1145/3687052

Haraway D (2008) Companion species, mis-recognition, and queer worlding. Queering the non/human pp xxiii–xxxvi

Hoffman MD, Blei DM, Wang C, et al (2013) Stochastic variational inference. J Mach Learn Res 14(1):1303–1347

Hooks B (2000) All about love: New visions

Horvitz Z, Chen J, Aditya R, et al (2024) Getting serious about humor: Crafting humor datasets with unfunny large language models. In: Ku LW, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 855–869, 10.18653/v1/2024.acl-short.76, URL https://aclanthology.org/2024.acl-short.76/

Johnson MK (2020) Joy: a review of the literature and suggestions for future directions. The Journal of Positive Psychology 15(1):5–24. 10.1080/17439760.2019.1685581, URL https://doi.org/10.1080/17439760.2019.1685581, https://doi.org/10.1080/17439760.2019.1685581

Kazim E, Koshiyama AS (2021) A high-level overview of ai ethics. Patterns 2(9):100314. https://doi.org/10.1016/j.patter.2021.100314, URL https://www.sciencedirect.com/science/article/pii/S2666389921001574

Kim S, Chilton L (2025) Ai humor generation: Cognitive, social and creative skills for effective humor. 10.48550/arXiv.2502.07981

Klein SE (2018) The value of play and the good human life. Cultura CCD 13(38):119–125

Kleiner S, Grieser JA, Miller S, et al (2025) Unmasking camouflage: exploring the challenges of large language models in deciphering african american language & online performativity. AI and Ethics 5(1):29–37

Lee A, Kummerfeld JK, Ann L, et al (2024) A comparative multidimensional analysis of empathetic systems. In: Graham Y, Purver M (eds) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, St. Julian's, Malta, pp 179–189, 10.18653/v1/2024.eacl-long.11, URL https://aclanthology.org/2024.eacl-long.11/

Lorde A (1988) A burst of light: And other essays. Courier Dover Publications, Mineola

Lu JH, Steele CK (2019) 'joy is resistance': cross-platform resilience and (re)invention of black oral culture online. Information, Communication & Society 22(6):823–837. 10.1080/1369118X.2019.1575449, URL https://doi.org/10.1080/1369118X.2019.1575449, https://doi.org/10.1080/1369118X.2019.1575449

Manzini A, Keeling G, Marchal N, et al (2024) Should users trust advanced ai assistants? justified trust as a function of competence and alignment. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '24, p 1174–1186, 10.1145/3630106.3658964, URL https://doi.org/10.1145/3630106.3658964

Maragh-Lloyd R (2021) From permit patty to karen: Black online humor as play and resistance. American Journal of Play 13:253–277

Mathewes C (2015) Toward a theology of joy. Joy and human flourishing: Essays on theology, culture, and the good life pp 63–96

Mihalcea R, Strapparava C (2005) Making computers laugh: Investigations in automatic humor recognition. In: Mooney R, Brew C, Chien LF, et al (eds) Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Vancouver, British Columbia, Canada, pp 531–538, URL https://aclanthology.org/H05-1067/

Mishra P, Rastogi C, Pfohl SR, et al (2025) Decoding safety feedback from diverse raters: A data-driven lens on responsiveness to severity. URL https://arxiv.org/abs/2503.05609, arXiv:2503.05609

Muwwakkil J, Hercules DM (in press) "slow it down, i just dissed you!": Musicolinguistic embodiment in battle rap. International Journal for the Sociology of Language

Patton DU, Kleiner S, Miller S, et al (2025) Digital narratives of grief and resilience: Insights from the integrating emotional stories online (ieso) platform. In: de la Iglesia DH, de Paz Santana JF, López Rivero AJ (eds) New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence. Springer Nature Switzerland, Cham, pp 368–378

Patulny R, Lazarevic N, Smith V (2020) 'once more, with feeling,' said the robot: Ai, the end of work and the rise of emotional economies. Emotions and Society 2(1):79 – 97. 10 .1332/263168919X15750193136130, URL https://bristoluniversitypressdigital.com/view/journals/emsoc/2/1/article-p79.xml

Pohlmeyer A, Desmet P (2017) From good to the greater good, Routledge, Abingdon; New York, pp 469–486

Rahman J (2007) An ay for an ah: Language of survival in african american narrative comedy. American Speech 82(1):65–96. 10.1215/00031283-2007-003, URL https://doi.org/10.1215/00031283-2007-003, https://read.dukeupress.edu/american-speech/article-pdf/82/1/65/395099/ASp82.1.3Rahman.pdf

Ramteke PS, Khandelwal S (2023) Comparing conventional machine learning and large-language models for human stress detection using social media posts. In: 2023 2nd International Conference on Futuristic Technologies (INCOFT), pp 1–8, 10.1109/INCOFT60753.2023.10425133

Register Y (2024) The future of ai can be kind: Strategies for embedded ethics in ai education. PhD thesis, University of Washington

Register Y, Tan Garcia JW, Kaushal N, et al (2022) Ai education matters: Guiding our future ai leaders with joy and justice. AI Matters 8(2):22–24. 10.1145/3557785.3557790, URL https://doi.org/10.1145/3557785.3557790

Rickford JR, Rickford RJ (2000) Spoken soul: The story of black English. Wiley, New York

Robertson RS, George J, Kuan Johnson M (2023) Building for joy in the digital world

Roose K (2024) Can a.i. be blamed for a teen's suicide? The New York Times

Sharma M, Tong M, Korbak T, et al (2024) Towards understanding sycophancy in language models. In: The Twelfth International Conference on Learning Representations, URL https://openreview.net/forum?id=tvhaxkMKAn

Siyan L, Shao T, Yu Z, et al (2024) EDEN: Empathetic dialogues for English learning. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 3492–3511, 10.18653/v1/2024.findings-emnlp.200, URL https://aclanthology.org/2024.findings-emnlp.200/

Slicer D (2015) More joy. Ethics and the Environment 20(2):1–23. URL http://www.jstor.org/stable/10.2979/ethicsenviro.20.2.1

Snyder CR, Lopez SJ (2001) Handbook of Positive Psychology. Oxford University Press, New York, 10.1093/oso/9780195135336.001.0001, URL https://doi.org/10.1093/oso/9780195135336.001.0001

Srivastava A, Sutton C (2017) Autoencoding variational inference for topic models. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=BybtVK9lg

Stevens R, Gilliard-Matthews S, Dunaev J, et al (2016) The digital hood: Social media use among youth in disadvantaged neighborhoods. New Media & Society 19(6):950–967. 10.1177/1461444815625941, URL http://dx.doi.org/10.1177/1461444815625941

Taylor Z (2022) "(it) shouldn't be funny but you can't help but laugh": Black twitter, #tweetlikethe1600s, and black humor online1. Social Media + Society 8(2):20563051221107631. 10.1177/20563051221107631, URL https://doi.org/10.1177/20563051221107631, https://doi.org/10.1177/20563051221107631

Tennen H, Affleck G (2002) Benefit-finding and benefit-reminding, vol 1, Oxford University Press, New York, pp 584–597

Turcan E, Muresan S, McKeown K (2021) Emotion-infused models for explainable psychological stress detection. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 2895–2909, 10.18653/v1/2021.naacl-main.230, URL https://aclanthology.org/2021.naacl-main.230/

Vanderlyn L, Weber G, Neumann M, et al (2021) "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In: Bisazza A, Abend O (eds) Proceedings of the 25th Conference on Computational Natural Language Learning. Association for Computational Linguistics, Online, pp 44–57, 10.18653/v1/2021.conll-1.4, URL https://aclanthology.org/2021.conll-1.4/

Wang B, Yue X, Sun H (2023) Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 11865–11881,

10.18653/v1/2023.findings-emnlp.795, URL https://aclanthology.org/2023.findings-emnlp.795/

Wang Y, Inkpen D, Kirinde Gamaarachchige P (2024) Explainable depression detection using large language models on social media data. In: Yates A, Desmet B, Prud'hommeaux E, et al (eds) Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024). Association for Computational Linguistics, St. Julians, Malta, pp 108–126, URL https://aclanthology.org/2024.clpsych-1.8/

Zargham N, Avanesi V, Reicherts L, et al (2023) "funny how?" a serious look at humor in conversational agents. In: Proceedings of the 5th International Conference on Conversational User Interfaces. Association for Computing Machinery, New York, NY, USA, CUI '23, 10.1145/3571884.3603761, URL https://doi.org/10.1145/3571884.3603761

THE PARIS JOURNAL ON AI & DIGITAL ETHICS

The social, ethical, and political implications of Artificial Intelligence and digital technologies have moved to the centre of public debate. Yet, these debates too often unfold without involving the relevant experts. As a result, a persistent gap remains between how public and private decision-makers perceive the ethical challenges of digital transformation and how these are understood by academic researchers and industry practitioners.

The Paris Conference on AI & Digital Ethics was founded to bridge this divide. It serves as a cross-disciplinary and cross-sectoral forum bringing together scholars from the humanities, social sciences, and the cognitive and computer sciences, alongside actors from industry, civil society, and politics. Within this space of exchange, participants are invited to confront methods, share perspectives, and collectively reflect on how technological innovation reshapes societies.

The Journal brings together the scholarly works presented at the Conference, developing them into full papers through double-blind review. By uniting critical insight with constructive engagement, it provides a forum where diverse disciplines may converge to examine the social transformations driven by AI and digital technologies—and to guide their development towards the common good.

The Paris Conference on AI & Digital Ethics

3 rue La Boétie, 75 008 Paris - Website: https://www.paris-conference.com

November 2025